A Comparison of Two Balance Calibration Model Building Methods

Richard DeLoach [1]
NASA Langley Research Center, Hampton, Virginia 23661

and

Norbert Ulbrich [2]
Jacobs Technology, Inc., Moffett Field, California

Simulated strain-gage balance calibration data is used to compare the accuracy of two balance calibration model building methods for different noise environments and calibration experiment designs. The first building method obtains a math model for the analysis of balance calibration data after applying a candidate math model search algorithm to the calibration data set. The second building method uses stepwise regression analysis in order to construct a model for the analysis. Four balance calibration data sets were simulated in order to compare the accuracy of the two math model building methods. The simulated data sets were prepared using the traditional One Factor At a Time (OFAT) technique and the Modern Design of Experiments (MDOE) approach. Random and systematic errors were introduced in the simulated calibration data sets in order to study their influence on the math model building methods. Residuals of the fitted calibration responses and other statistical metrics were compared in order to evaluate the calibration models developed with different combinations of noise environment, experiment design, and model building method. Overall, predicted math models and residuals of both math model building methods show very good agreement. Significant differences in model quality were attributable to noise environment, experiment design, and their interaction. Generally, the addition of systematic error significantly degraded the quality of calibration models developed from OFAT data by either method, but MDOE experiment designs were more robust with respect to the introduction of a systematic component of the unexplained variance.

Nomenclature

AF	=	axial force
d	=	order of polynomial
F	=	load symbol
i, j	=	index variables
k	=	number of independent variables
K	=	number of regressors in a polynomial model
NF	=	normal force
p	=	number of parameters in a polynomial model, including intercept
PM	=	pitching moment
rAF	=	axial force gage response
rNF	=	normal force gage response
rPM	=	pitching moment gage response
RM	=	rolling moment
rRM	=	rolling moment gage response
rSF	=	side force gage response

[1] Senior Research Scientist, Aeronautical Systems Engineering Branch, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Senior Member.
[2] Aerodynamicist, Jacobs Technology, Inc., NASA Ames Research Center.

rYM	=	yawing moment gage response
SF	=	side force
X	=	design matrix
YM	=	yawing moment

Terminology

Alternative Hypothesis	=	Assertion that an effect is non-zero by a significant amount
ANOVA	=	Analysis of Variance
Bonferroni Limt	=	Minimum acceptable probability that multiple terms are significant
Coding	=	Linear transformation of variables into a range convenient for processing
Confidence Interval	=	Precision Interval when n = infinity
Explained SS	=	Sum of Squares attributable to known causes
F-Value	=	Ratio of Mean Square for an effect to Residual mean square
Hierarchy	=	Condition in which higher order terms are accompanied by component lower-order term
Inference	=	Decision to reject either a null hypothesis or its corresponding alternative
Inference Space	=	A coordinate system in which one axis is assigned to each independent variable
Interaction Effect	=	Change in effect due to change in factor level from low to high
LOF	=	Lack of Fit
LSD	=	Least Significant Difference
Main Effect	=	Change in response due to change in factor level from low to high
MDOE	=	Modern Design of Experiments
Mean Square	=	Ratio of sum of squares to degrees of freedom. Variance
Multicollinear	=	State in which two or more regressors share a near-linear dependency
Normal Probability Plot	=	Graph distinguishing between random and systematic effects
Null Hypothesis	=	Assertion than an effect is zero
OFAT	=	One Factor At a Time
Orthogonal	=	State in which regressors are all mutually independent
Pareto Chart	=	Bar chart of ordered absolute t-Values
Precision Interval	=	Range in which the average of an n-point sample is expected to lie a prescribed percentage of the time
Prediction Interval	=	Precision Interval when n = 1
PRESS	=	Predicted Residual Sum of Squares
Regressor	=	Term in a regression model
Residual	=	Difference between measurement and some reference
Residual Mean Square	=	Residual sum of squares divided by residual degrees of freedom
Residual SS	=	Difference between total sum of squares and explained sum of squares
Significance	=	Risk of erroneously rejecting a null hypothesis
SVS	=	Single Vector System
t-Limit	=	Minimum acceptable probability that a given term is significant
t-Value	=	Measured quantity expressed as multiple of standard error in measurement
Total SS	=	Total sum of squares. Measure of variability in data set
VIF	=	Variance Inflation Factor/ A measure of multicollinearity

I. Introduction

A force balance is used in a typical wind tunnel test to measure aerodynamic load components by generating electrical signals that are proportional to the strain induced by those loads. The relationships between applied load, induced strain, and resulting electrical signal are complex; while ideally each balance signal would represent only its component of the applied load and have no response to any other load components, in practice the electrical outputs are influenced by multiple load interactions. Balance designs are optimized to minimize these undesirable interactions, but in practice they cannot be entirely eliminated. For this reason, it is necessary to carefully calibrate force balances.

There have been many changes in the design of force balance transducers since the first application of measured strain to infer aerodynamic loads in the 1940s, including the development of machining methods required to produce single-piece, six-component balance designs. But apart from automating elements of the force balance calibration

process (implemented by some but not all laboratories), there have been relatively few changes in basic force balance calibration methods in the ensuing 60 years.

The goal of a force balance calibration experiment is to derive a mathematical model that defines the relationship between the loads applied to the balance and the electrical signals it produces. The accuracy of this model is also determined during the calibration experiment. It is a common convention in the aerospace force measurement community, as well as a formal recommendation of the American Institute of Aeronautics and Astronautics,[1] to use a math model in the form of a polynomial function of the component loads. Quoting from the AIAA's Recommended Practice on Calibration and Use of Internal Strain-Gage Balances with Application to Wind Tunnel Testing, "In its most common form, the model assumes the electrical output reading from the strain-gage bridge for the i^{th} component (R_i) to be related to the applied single and two-component loads … by a second order polynomial function…"

This convention of using second-order polynomials as the calibration model appears to be based on early results that argue against the need for higher-order terms. For example, the AIAA Recommended Practice cites a 1959 report by the British Royal Aircraft Establishment,[2] which states, "In practice, for a well-designed balance, it is generally found that terms of the third and higher degrees in load components are completely negligible, while coefficients of second degree terms (called second order coefficients) are nearly all small, if not negligible…" An even earlier (1956) report of the National Advisory Committee for Aeronautics (forerunner of NASA) states,[3] "To date the Langley Laboratory has not encountered any terms for these equations which are higher than the second order, although many load combinations are made which would reveal many third-order terms if they were present." The AIAA Recommended Practice document notes that the second-order polynomial model is sometimes extended by the addition of pure cubic terms (although mixed cubic terms—third-order terms of the form A^2B—are not mentioned). The document also describes the addition of certain absolute-value terms to account for situations in which the outputs of a balance depend on the sign of the strain in the measuring elements, as is not uncommon in multi-piece balance designs.

Whether the basic second-order polynomial model or one of the extensions to this base model is used, we can assume that in a calibration experiment there will be K regressors in the math model and that a load schedule consisting of n combinations of component loads will be planned with n corresponding response measurements, where n must be greater than $K + 1$ for all regression coefficients to be estimated including the intercept. The general form of the full polynomial model is then as follows:

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i \tag{1a}$$

where y_i is the response recorded for the i^{th} load application, x_{ij} is the i^{th} level of the j^{th} regressor, β_j is the coefficient of the j^{th} regressor, and ε_i is an error term, assumed to be drawn from a normal distribution with a mean of 0 and with a constant standard deviation for all responses. The quantity β_0 is an intercept term that requires certain additional iterative procedures to account for the fact that an unloaded balance still experiences applied forces due to the weight of the loading hardware and the balance itself. These procedures are beyond the scope of the current paper but are addressed in detail in the AIAA Recommended Practice on balance calibration cited above.

Equation (1a) can be described more succinctly in vector/matrix form as follows:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1b}$$

where $\mathbf{y}$ is an $(n \times 1)$ vector of response measurements, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of coefficients, and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of error terms.

$\mathbf{X}$ is called the design matrix, consisting of n rows corresponding to the number of data points acquired in the experiment, and p columns, one for each term in the math model, including the intercept term. The design matrix plays a central role in determining both the cost and the quality of a calibration experiment,[4] and indeed of any experiment in which a polynomial math model is used to fit system responses as a function of multiple independent variables. Furthermore, the details of the design matrix are largely within the researcher's control. The number of rows depends on the volume of data that the researcher decides to acquire, the number of columns depends on the math model that the researcher selects, and the values of the individual matrix elements depend on the levels of the independent variables that are set for each data point. For a balance calibration experiment, this is defined by what the researcher specifies as the load schedule.

We focus on certain elements of the design matrix in this paper because it is such an important determinant of both productivity and quality in a response modeling experiment, and because it is so easy for the researcher to control. It is obvious that productivity can be influenced by the volume of data acquired, which is defined by the number of rows in the design matrix. Direct operating costs and cycle time can both be minimized by specifying as few rows in the design matrix as will be adequate to achieve the objectives of the experiment.[5] The impact of the design matrix on quality is revealed through the covariance matrix and how it affects uncertainty both in estimates of the individual regression coefficients, and in response predictions made by applying the model.

The covariance matrix, **C**, is a ($p \times p$) square matrix computed by pre-multiplying the design matrix by its transpose, inverting the product, and multiplying each element of the resulting matrix by the unexplained variance of the residuals, σ^2:

$$\mathbf{C} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2 \tag{2}$$

It can be shown that the diagonal elements of the covariance matrix represent the variance in estimates of the regression coefficients. That is, the variance in the i^{th} regression coefficient, β_i is simply:

$$Var\left(\beta_i\right) = \mathbf{C}_{ii} \tag{3}$$

The off-diagonal elements of the covariance matrix quantify the degree to which the regressors are correlated. Correlated regressors result in an undesirable characteristic known as multicollinearity, of which more will be said presently. We would generally prefer that the off-diagonal elements of the covariance matrix all be zero. While this ideal state is difficult to achieve in all practical circumstances, much of the quality improvement delivered by formally designed experiments is derived by defining the test matrix in such a way as to minimize the off-diagonal elements of the covariance matrix. The impact of formal experiment design in a balance calibration experiment will be discussed further in later segments of this paper.

Other than the intrinsic variability of the data itself, Eqs. (2) and (3) clearly indicate that the variance of the estimated regression coefficients depends *only* on the design matrix! This suggests that there may be opportunities to minimize the uncertainty in estimates of the regression coefficients simply by optimizing the design matrix in some way.

Similarly, when the model is used to predict responses for a specified combination of independent variable settings (component loads in a balance calibration experiment), the variance in those predictions depends heavily on the design matrix. Consider a vector $\mathbf{x}_0 = [1\ x_{01}\ x_{02}\ \ldots\ x_{0K}]'$ representing a data point specified by a given combination of component loads on a balance, where x_{01} is the level of the i^{th} regressor corresponding to this point. The estimated mean response at this point is

$$\hat{\mathbf{y}}\left(\mathbf{x_0}\right) = \mathbf{x_0'b} \tag{4}$$

where **b** is a vector of estimated regression coefficients. That is, **b** is our best estimate (typically by some least-squares criterion) of the vector of true coefficients, β, in Eq. (1b). The set of estimated mean responses over all $\mathbf{x}_0$ comprise what is called a response surface, and the process of estimating the **b** vectors is called response surface modeling. The variance in the response prediction at a particular $\mathbf{x}_0$ is computed as follows:

$$\mathrm{Var}\left[\hat{\mathbf{y}}\left(\mathbf{x}_0\right)\right] = \sigma^2\mathbf{x_0'}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x_0} \tag{5}$$

As in the case of the individual regression coefficients, Eq. (5) reveals that except for the intrinsic variability in the data used to fit the response model, the variance in a response prediction for a given point is determined entirely by the design matrix. Equations (2), (3), and (5) suggest that since the design matrix plays such a central role in determining the precision of a response surface modeling experiment, improvements in the quality of such experiments can be achieved by planning the design matrix carefully.

Research has been ongoing at NASA Langley Research Center for about 10 years to demonstrate the practical improvements that can be achieved in aerospace research quality and productivity generally, by optimizing the design matrix in multivariable response modeling experiments. This research has been an element of a larger testing technology activity geared toward enhancing aerospace research quality and productivity by integrating experiment design, execution, and analysis activities into a formal research process described as the Modern Design of Experiments (MDOE). The MDOE process has been successfully applied in numerous aerospace research disciplines at Langley and elsewhere. This includes the area of balance calibration, where an entirely new hardware system was introduced to facilitate the MDOE loading combinations needed to optimize the design matrix while retaining the simplicity and inherent reliability of the traditional dead-weight loading method preferred at Langley.[6,7] It should be noted that MDOE design matrix optimization is also well-suited for automated balance calibration machines.[8,9] Such machines are capable of applying multiple loading combinations without certain physical constraints inherent in a traditional dead-weight calibration system, although at the expense of some additional complexity in alignment.[10]

Independent of the MDOE activity at Langley, NASA Ames Research Center has investigated a key aspect of the design matrix optimization problem with respect to balance calibration; namely, the selection of math models that can be supported by a given load schedule and that also minimize the residual variance when the model is fitted to the calibration data. In the broader context of the design matrix, the Ames effort focuses on the columns of that matrix, identifying a recommended subset of the largest math model that can be supported by a given load schedule. To motivate subsequent discussions in this paper of practical implementation details, we now present a brief general explanation for why the quality of a response surface modeling experiment can often be improved by eliminating some of the math model terms.

Assume for a moment that Eq. (1) represents the largest math model that can be supported by a given test matrix. For a balance calibration experiment, the test matrix is simply the calibration load schedule. That is, we assume that this equation describes the largest math model for which non-singular regression results can be obtained for the prescribed load schedule, subject to the constraint that only model terms drawn from a prescribed original set are considered. This constraint simply ensures that the order of the model and the functional form of the individual regressors is consistent with standard recommended practices, as described above.

We wish to examine the consequences of reducing the model so that fewer than the K regressors of the model described by Eq. (1) are retained. Let r represent the number of regressors that we wish to remove and let p represent the number of terms that will be retained in the model, including the intercept term. We can then express Eq. (1) as follows:

$$\mathbf{y} = \mathbf{X}_p \beta_p + \mathbf{X}_r \beta_r + \varepsilon \tag{6}$$

Here, $\mathbf{X}_\mathrm{p}$ is a ($p \times n$) matrix with columns corresponding to the retained terms in the model, including the intercept, and β_p is a ($1 \times p$) vector of the corresponding regression coefficients for this reduced model. The columns of $\mathbf{X}_\mathrm{r}$ represent terms that are deleted from the model, and β_r is a vector of the corresponding regression coefficients.

If $\mathbf{b}$ is a vector of estimated regression coefficients for the unreduced model, and $\mathbf{b}_\mathrm{p}$ corresponds to those coefficients that are retained, it can be shown that the matrix $\mathrm{Var}(\mathbf{b}_\mathrm{p}) - \mathrm{Var}(\beta_\mathrm{p})$ is positive semidefinite.[11] Therefore, dropping terms from the full model and refitting the data to a subset of the original regressors results in model coefficient estimates with variance that is less than or equal to the variance in the corresponding coefficients of the full model. In other words, with respect to the precision of the regression coefficient estimates there is nothing to lose, and possibly something to gain, by reducing the number of regressors in the math model.

Consider now the impact of such a model reduction on the variance of response predictions. Note that a vector of predicted responses for each point in the test matrix can be generated from the vector of measured responses, $\mathbf{y}$, by means of the "hat matrix," $\mathbf{H}$, as follows:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \tag{7}$$

where

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \tag{8}$$

and **X** is the design matrix, as before. The variance in the vector of response estimates is computed as follows:

$$\mathrm{Var}(\hat{\mathbf{y}}) = \mathbf{H}'\mathrm{Var}(\mathbf{y})\mathbf{H} = \mathbf{H}'(\mathbf{I}\sigma^2)\mathbf{H} \tag{9}$$

The hat matrix is both symmetric (equal to its transpose) and idempotent, meaning that $\mathbf{HH} = \mathbf{H}$. Equation (9) therefore reduces to

$$\mathrm{Var}(\hat{\mathbf{y}}) = \mathbf{H}\sigma^2 \tag{10}$$

Note that the variance of the i^{th} response prediction is just the i^{th} diagonal element of $\mathbf{H}\sigma^2$. Following Box and Draper,[12] we consider the trace of this matrix, which is just the sum of all the diagonal elements:

$$\mathrm{trace}(\mathbf{H}\sigma^2) = \sigma^2\mathrm{trace}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] = \sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i) \tag{11}$$

We invoke the following matrix identity: trace(**AB**) = trace(**BA**). Let **A** = **X** and **B** = $(\mathbf{X'X})^{-1}\mathbf{X'}$. Then

$$\mathrm{trace}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] = \mathrm{trace}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right] = \mathrm{trace}(\mathbf{I}_p) = p \tag{12}$$

since p is the dimension of the square matrix $(\mathbf{X'X})^{-1}$ as noted above in the description of the covariance matrix. Combining Eqs. (11) and (12) we have

$$\sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i) = p\sigma^2 \rightarrow \frac{\sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i)}{n} = \frac{p\sigma^2}{n} \tag{13}$$

That is, the prediction variance averaged over all points in a regression analysis is the same for any order model and depends only on the intrinsic variability of the data, the volume of data fitted, and the number of terms in the math model. Therefore, if a given number of data points are acquired in a measurement environment with a given degree of intrinsic variability, the average prediction variance depends on nothing more than the number of terms in the regression model! Obviously, there is potential to reduce the average prediction variance simply by reducing the number of terms in the model. This is because each term in the model carries with it some contribution to the total uncertainty in the prediction.

The purpose of this introduction has been to identify certain benefits of reducing the number of terms in a response model, to motivate the examination of practical means of implementing such reductions which comprises the bulk of this paper. Summarizing the key points, the design matrix determines the uncertainty in estimates of the regression coefficients and the model predictions. Reducing the number of columns in the design matrix by selecting a reduced math model never increases the variance in estimates of the retained coefficients, and may decrease it. Because the prediction variance averaged over all regression points is directly proportional to the number of terms in the math model, reducing the number of model terms also increases the precision of the response predictions.

The remainder of this paper describes practical ways and means of improving the quality of a balance calibration by reducing the calibration math model. These techniques utilize statistical metrics to evaluate the significance of math model terms that may be used for the global regression of calibration data. Two related methods are compared in this paper.

The first method is called “Candidate Math Model Search.” This technique was first developed at Ames Research Center for the analysis of strain-gage balance calibration data.[13,14] The method has been implemented in a software package called BALFIT.[15]

The second method is currently being used at Langley Research Center for balance calibration analysis and other data analysis problems. It relies upon standard stepwise regression analysis methods[11,12,16] that are implemented in a number of commercially available software packages.[17-21] Table 1 summarizes key elements of the two math model

building methods. The specific stepwise regression procedures of a representative commercial package (Design Expert® version 7, or DX7) are compared with the BALFIT variable selection and model building process.

Table 1. Description of basic elements of two balance calibration math model building methods.

SOFTWARE TOOL	DESCRIPTION
BALFIT	First, Singular Value Decomposition (SVD) is used in order to get the largest math model, i.e., the permitted math model, that the calibration data supports; then, using the permitted math model as an upper bound, candidate math models are found using the standard deviation of the response residuals as a metric; finally, the recommended math model is assembled from the candidate math models using a user selected change of the standard deviation from one math model to the next as a metric.
DX7	The independent variables are specified and upper and lower limits are defined to facilitate coding of variables. Factor interaction models and all full d^{th}-order polynomial models (d from 1 to 6) are automatically examined to identify a subset of permitted models (full-rank hat matrix so nonsingular) that the test matrix will support. A weighted combination of residual standard deviations, PRESS statistics, lack-of-fit F-statistics, and R-squared statistics is used to suggest an initial full model. The user can edit this or select another full permitted model as the starting point for the variable selection and model building process. Terms from the starting model are discarded or retained by the stepwise regression algorithm, which seeks to minimize unexplained variance and multicollinearity. Hierarchy is imposed on the resulting reduced model at the user's option. Multiple metrics for fit quality are tabulated and displayed graphically. The user edits the final recommended model based on these metrics to incrementally improve the fit, to resolve multicollinearity issues, or to reflect the user's subject matter expertise.

Each of the two methods determines an optimum math model for the analysis of balance calibration data that depends on the calibration data set. In addition, both methods still require some user input in order to uniquely specify the optimum math model. Table 2 below lists the user input that is needed.

Table 2. User input needed to specify optimum math model.

SOFTWARE TOOL	USER INPUT
BALFIT	• Combination of math term groups, i.e., function classes, that are used for the global regression of the balance calibration data (see Table 7; at the present time the BALFIT software supports all math term groups) • Standard deviation change from one candidate math model to the next in percent of the standard deviation minimum of each gage
DX7	• Input variables and ranges • Alpha-in and Alpha-out: Significance levels for retaining and rejecting regressors from the permitted math model • Subject matter expertise to assess quality of recommended model and to edit/reassess as necessary

Both math model building methods are being applied to real-world data analysis problems on a regular basis. However, so far no direct comparison of the two methods has been performed in order (i) to compare selected sets of math terms, and (ii) to assess residuals of fitted data sets. Recently, the first author suggested that numerically simulated balance calibration data be used to compare the two methods. This approach has several advantages: (i) the exact solution of the global regression problem of the simulated data sets is known; and (ii) random and systematic errors can easily be introduced in the data sets in order to investigate the impact of data errors on the math model building methods.

In the first part of this paper, the generation of the four simulated balance calibration data sets is discussed. Then, key elements of the two math model building methods are reviewed. Finally, results of the application of both math model building methods to the simulated calibration data sets are presented.

II. Balance Calibration Data Simulation

A reduced second-order polynomial function of six variables was used to simulate a six-component strain gage balance. For realism, the significant interaction terms and the relative magnitudes of the coefficients were modeled loosely after a production wind tunnel balance in current use at Langley Research Center, although the model as simulated does not precisely represent any specific balance at Langley or elsewhere. Table 3 lists the significant terms and coefficients for the simulated model and Table 4 displays the simulated load capacities. Note that the model coefficients correspond to independent variables in coded units, not physical units. The coded variables represent a linear transformation of calibration loads into a range from -1 to +1, where -1 corresponds to the algebraically smallest load capacity while +1 corresponds to the largest load capacity. Table 4 displays the correspondence between range limits for coded and physical units. Commercial data analysis packages such as the one used to create the model for the simulated balance tend to use coded variables, for reasons outlined in Appendix A.

Table 3. Coefficients of "true balance," coded variables.

	Intercept	A-NF	B-AF	C-PM	D-RM	E-YM	F-SF	A^2	B^2	C^2	D^2	E^2	F^2	AB	AC	AD	AE	AF	BC	BD	BE	BF	CD	CE	CF	DE	DF	EF
NF	-1.0	**2096.7**	-1.0	5.1	-1.4	-0.1	-0.2	2.16		0.75	-0.88		-0.78	0.26					-0.56						-1.82	-0.49	-7.08	0.69
AF	538.6	3.0	**538.7**	-2.6	5.0	-2.0	1.0	-1.13		0.40	0.30	-0.35	-0.30	0.30		0.82			-0.75							-1.20	-4.00	0.30
PM	-0.1	6.6	-0.1	**1397.6**	9.1	3.4	0.9	0.79			-0.45	0.43	0.69		1.14	-0.43			-0.69				0.00		-0.89	-2.58	-8.95	
RM	0.0	-1.5	-0.1	1.9	**592.9**	1.0	-1.0	0.33		0.26	-0.14	0.38				-0.40			0.21				1.17		0.40			
YM	-0.2	11.4	-0.1	-3.5	12.9	**1094.6**	-5.7	-0.84			-0.32	0.34				16.49					-0.44		10.12				0.62	
SF	0.0	-3.5	-0.1	-0.7	-12.4	-4.1	**1974.7**	0.40			0.21		-0.54	-0.86		17.13		-8.41	0.00		-0.43	-0.51	0.00		1.96		1.01	0.00

Two load schedules were simulated. The first was a conventional calibration load schedule of the kind that has been used routinely at Langley Research Center for decades and is similar in data volume and general characteristics to calibration sequences employed at numerous other laboratories. It was originally designed to permit model coefficients to be determined by graphical means, reflecting a computational limitation of the 1950s when it was initially introduced. The basic design has remained virtually unchanged since then. There are a total of 81 load sequences performed sequentially in time. Each load sequence consists of a tare point, four increments, three decrements, and a return tare point providing a total of nine data points per sequence; 729 points in all. This is a member of a general class of test designs known in the testing technology literature as One Factor At a Time (OFAT) designs, reflecting the fact that in each sequence only one component is loaded at a time while the others are held constant.

The second load schedule is an MDOE design for which the volume of data has been scaled to achieve specified precision goals in a given measurement environment, and the loading combinations have been selected to optimize the design matrix in certain ways. The specific details of the MDOE experiment design process are beyond the scope of this paper, but we mention some of the principal design features:

1) There were 64 points in the MDOE design, compared to 729 in the OFAT design.
2) The component loads were selected to minimize off-diagonal elements of the covariance matrix, maximizing prediction precision for a given volume of data.
3) The points were selected to facilitate orthogonal blocking, a type of point-grouping that allows between-group bias errors to be conveniently removed from the regression computations. Blocking also facilitates the estimation of bias errors induced by systematic variations.
4) The points were executed in random order, a standard MDOE quality assurance tactic by which systematic variations washing through the calibration facility over time (thermal effects, instrument drift, etc.) are prevented from biasing the estimation of regression coefficients other than the intercept. These systematic bias errors are converted to random fluctuations that can be compensated by acquiring additional data if necessary.

A third set of simulated data consisted of 25 component load combinations that were randomly selected from a uniform distribution of values ranging between the load capacity limits of the balance as displayed in Table 4. These points were not used to fit any of the calibration response models. Instead, they were held in reserve as "confirmation points" to be used to test the models, as will be described later in the paper.

The models in Table 3 represent the "true" balance models in this simulation, as distinct from models to be determined later by fitting simulated data. A perfect regression analysis would simply return these models exactly, and we use departures in the fitted models from these true response models to assess the regression model-building process.

Table 4. Load limits (pounds for forces, inch-pounds for moments) and corresponding coded values.

Coded	NF	AF	PM	RM	YM	SF
-1	-6520	0	-12800	-8150	-6400	-4000
+1	6520	400	12800	8150	6400	4000

Simulated error-free balance responses were generated for each of the two calibration load schedules as well as for the 25 simulated confirmation points, using the coefficients in Table 3. Two error environments were then simulated. For the first error environment, simulated experimental errors were drawn at random from a normal distribution with a mean of zero and a constant standard deviation, using ordinary Monte Carlo techniques. The standard deviation of these simulated random errors was determined from 324 measured replicates in an actual 729-point OFAT balance calibration conducted at Langley Research Center using a balance with the same load capacity as the simulated balance in this study. Table 5 lists the standard deviations of these replicated responses and Fig. 1a shows the actual simulated random errors normalized by standard deviation. Note that the normalized simulated random errors in Fig. 1a are centered on zero and approximately 95% of them lie within $\pm 2\sigma$, confirming the designed distributional features of the random error simulation.

Table 5. Standard deviation of random error, MicroV/V.

NF	AF	PM	RM	YM	SF
0.50	0.61	0.63	0.20	1.04	1.30

The second simulated error environment features the same random noise as the first, but included an additional component of systematic error. Systematic errors are due to long-period, slowly varying effects that occur routinely in real-world experiments. These effects cause sample means to vary slowly over time, violating a common but unfortunately somewhat naïve assumption that all random errors occur about stable (time-invariant) sample means.

These time-dependent bias errors are caused by such factors as slowly changing temperature effects, instrumentation drift, operator fatigue, system warm-up effects, mechanical expansion or contraction due to temperature or applied loads, and any number of additional unknown (and unknowable) causes. In high-precision research environments such as are common in aerospace research, and especially in calibration experiments where chance variations in the data are truly small, these systematic errors often comprise that largest component of the unexplained variance in a data set.

To simulate systematic error, there are an infinite number of choices for how sample mean might vary with time. In this simulation we chose a simple first-order function of time in which the systematic bias error component starts at minus six standard deviations at the beginning of the calibration and increases at a uniform rate sufficient to bring it to plus six standard deviations over the period of time in which a 729-point, hand-loaded OFAT calibration load schedule might be executed (typically 3 to 4 weeks). Figure 1b shows the combination of random and systematic errors normalized by standard deviation. Compare with Fig. 1a to see the effect of systematic error.

Note that due to the high precision that is characteristic of a calibration laboratory, the standard deviation of random variations tends to be relatively small, so that a drift of ±6 standard deviations over 3 to 4 weeks does not represent an especially large change in absolute terms. Also note that real-world systematic error is seldom so accommodating as to vary uniformly in time and in one direction only. In actual experiments, abrupt shifts are not uncommon and complex systematic error patterns are the norm. In that sense this simulation is a fairly conservative representation of typical systematic error effects. Nonetheless it serves to contrast the case of random-only error with that in which some systematic error is also present.

Table 6. Description of simulated balance calibration data sets.

DATA SET	LOAD SCHEDULE	ERROR TYPES	NUMBER OF POINTS
1	OFAT	Random	729
2	OFAT	Random + Systematic	729
3	MDOE	Random	64
4	MDOE	Random + Systematic	64

There were thus four simulated data sets, comprised of all four combinations of the two test matrix designs and the two noise environments. Table 6 summarizes their characteristics.

III. Candidate Math Model Search

During the past two years a new method was developed at Ames Research Center that analyzes strain-gage balance calibration data. The method combines industry-wide accepted balance calibration analysis procedures with automatically generated math models. The new Ames approach became possible after it was recognized that ideas from vector algebra and a numerical technique called Singular Value Decomposition (SVD) may be used to rigorously identify the largest math model that the calibration load schedule supports.[15] The flowchart in Fig. 2 summarizes key elements of the new Ames approach.

Experience has shown that the new Ames approach is fast, reliable, and accurate. It was successfully applied to a wide variety of in-house and customer supplied balance calibration data sets.[22,23] The approach uses applied calibration loads, measured electrical responses, and natural zeros of the balance in combination with a math model determination algorithm in order to generate two math models for the analysis. Calibration load or check load residuals have to be compared in order to decide which one of the two math models should be used for the final analysis of the data and the calculation of the data reduction matrix.

The first math model is called the permitted math model. It is the result of applying SVD to the tare corrected calibration load schedule. It is the largest math model that the load schedule supports using the user's math term group selection as a constraint. The global regression problem of the calibration data would become singular or close to singular if only a single additional term from the selected math term group combination would be added to the permitted math model. Table 7 shows different math term group choices that are traditionally used for strain-gage balance calibration analysis.

Table 7. Traditional math term group choices for strain-gage balance calibration analysis.

GROUP NUMBER	MATH TERM $F \Longrightarrow$ *load symbol* $i, j \Longrightarrow$ *gage indices*	DEFAULT FOR SINGLE-PIECE BALANCE	DEFAULT FOR MULTI-PIECE BALANCE
1[†]	F_i	×	×
2	$\lvert F_i \rvert$	-	×
3	$F_i \cdot F_i$	×	×
4	$F_i \cdot \lvert F_i \rvert$	-	×
5	$F_i \cdot F_j$	×	×
6	$\lvert F_i \cdot F_j \rvert$	-	-
7	$F_i \cdot \lvert F_j \rvert$	-	-
8	$\lvert F_i \rvert \cdot F_j$	-	-
9	$F_i \cdot F_i \cdot F_i$	-	-
10	$\lvert F_i \cdot F_i \cdot F_i \rvert$	-	-

[†] This group is included in all allowed group combinations.

The BALFIT software allows the user to choose practically any combination of these math term groups as a constraint for the permitted math model. By default, however, terms from the first group are always included in the user's selection. Experience showed that a combination of terms from the first, third, and fifth math term group will lead to good math models for a single-piece balance. A combination of terms from the first five groups is usually needed for the global regression of calibration data of a multi-piece balance.

The second math model is called the recommended math model. It is the result of applying a candidate math model search algorithm to the calibration data set using the permitted math model as an upper bound. The standard deviation of the response residual of different math term combinations is used as a metric in order to test individual terms of a possible candidate math model for significance. Then, using the standard deviation change from one candidate math model to the next as a metric, the recommended math model is assembled from the candidate math models. By design, the recommended math model only uses the most significant terms of the permitted math model. Table 8 below lists definitions of the permitted and the recommended math model. Figure 3 summarizes key elements of the determination of the permitted and recommended math model.

Table 8. User input needed to specify optimum math model.

MATH MODEL	DEFINITION
PERMITTED	Largest math model that the calibration data supports using the user's math term group selection as a constraint; it is determined by applying SVD to the tare corrected load schedule; terms of the math model are identical for all gages; it is used as an upper bound during the candidate math model search.
RECOMMENDED	The math model that uses only the most significant terms of each gage; the model is built using candidate math models that were generated after applying a search algorithm to the calibration data set; this search algorithm uses the permitted math model as a starting point.

In the end, the standard deviation of the calibration load residuals and the largest load residual of each gage should be compared in order to decide if the permitted or the recommended math model should be used for the final analysis of the calibration data. In addition, the accuracy of each math model should be evaluated using check or proof loads that were *not* a part of the original calibration load schedule. It is important to remember that the recommended math model uses only the most significant terms of each gage. Therefore, the recommended math model should always be used for the final analysis of balance calibration data whenever "over-fitting" is a concern. The recommended math model is also compared in the present study with the math model that is the result of applying stepwise regression analysis to the calibration data.

The first simulated data set of the present study may be used to demonstrate the application of the Ames approach to calibration data. The first, third, and fifth math term group from Table 7 were selected for the analysis. Figure 4a shows the corresponding permitted math model for this data set. All possible load combinations were used in the calibration data set. Therefore, SVD correctly recognized that all terms of the selected math term group combination are supported. Figure 4b shows the calibration load residuals in percent of the gage capacity after the global regression of the calibration data was performed using the permitted math model. In the next step, the candidate math model search was performed using the permitted math model as an upper bound. Figure 4c shows the result of the candidate math model search, i.e., the standard deviation of the response residual as a function of the number of terms of each candidate math model. Enlarged symbols in Fig. 4c identify those candidate math models that were used to assemble the recommended math model. A user specified standard deviation change of 1% of the standard deviation minimum of each gage was used to assemble the recommended math model. Terms of the recommended math model are depicted in Fig. 4d. Calibration load residuals are depicted in Fig. 4e after the recommended math model was used for the analysis. As expected, these load residuals are very close to the residuals that are depicted in Fig. 4b for the much larger permitted math model.

In the next part of the paper, the second math model building method is explained in more detail.

IV. Analysis by Stepwise Procedures

This section describes the application of standard stepwise regression methods to improve the design matrix of a calibration experiment by selecting terms for the response model. As noted earlier, these methods are implemented in a number of commercial data analysis software packages. To illustrate the stepwise regression method, an analysis of Data Set 1 (Table 6) is presented, paralleling the BALFIT analysis described in the previous section. This analysis was performed using one of many commercially available software packages—Design Expert[17] (AKA DX7)—although other commercial software implementations of this method[18-21].use similar procedures.

The DX7 analysis process begins by naming the dependent (response) and independent (load) variables. Upper and lower limits of the ranges of the independent variables are entered in order to facilitate coding them by a linear transformation that centers and scales them into a range from -1 to +1. The coding transformations are described in Appendix A. All of the commercial response surface modeling software packages cited in the references invoke the coding of independent variables, for reasons summarized in Appendix A. For the current analysis, this step simply meant entering the data from Table 4 above, in response to prompts from the software. The test matrix and all response measurements were then entered, simply by pasting them from a spreadsheet into an equivalent structure within DX7.

Once the data are entered, the DX7 stepwise regression analysis proceeds through four phases for each response variable: Fit Summary, Model Selection, Analysis of Variance, and Diagnostic Evaluation. As part of the diagnostic evaluation, a test known as the Box-Cox Transformation Test is automatically performed to determine if a power

transformation of the responses would provide a better fit to the data. If so, a specific transformation is recommended.

For the present analysis, no transformations were recommended for any of the data sets, but there can be data sets for which a better fit can be achieved by modeling the log of the response, say, or the square root. (A class of transformations known as "variance stabilization transformations" serves to make the unexplained variance more constant, for example, which adheres more closely to basic assumptions in least-squares regression analysis and therefore often generates better models.) When a specific transformation is recommended, the user has the option of selecting it from several choices in a drop-down menu. This automatically performs the specified transformation of the data. It is possible to quickly analyze the transformed data to see if there is any significant improvement over the untransformed case. If not, one of the transformation selections is "none," which can be invoked to return to an analysis of the original, untransformed data.

Once the transformation decision is made for a given response, the four main phases of the analysis can begin.

A. Phase One: Fit Summary

Phase One is analogous to the BALFIT process for identifying the "Permitted Model." DX7 performs a preliminary analysis that begins with the most simple of models—a "0^{th}-order" model consisting of only the intercept term—and continues through models of progressively higher order. The highest-order model that DX7 considers is dictated by the number of independent variables, as there is an upper limit of 400 terms that DX7 can include in any one model. All regression analyses in DX7 apply standard QR Decomposition to the design matrix to compute model coefficients.

Each model is first evaluated to determine if its hat matrix (Eq. (8)) is of full rank so that the model can be estimated. Note that the hat matrix is a function only of the design matrix, which depends only on the load schedule and not any of the responses. Therefore, the same set of models will pass this screen for all responses. Models lacking a full-rank hat matrix are flagged in red and a warning is provided not to select them. See Fig 5a, the Sequential Sum of Squares Report generated by the analysis of normal force responses for Data Set 1. This table provides a sequential comparison of models showing the statistical significance of adding more model terms to those already in the model. The significance can be assessed by examining the right-most column, labeled "p-value." This is the probability that an improvement due to extending the model represents an effect so small that it cannot be distinguished from noise. So, for example, the line labeled "Linear vs. Mean" has a p-value < 0.0001. This means that for this set of Normal Force data, a model consisting only of the six first-order terms plus the intercept is such a great improvement over a model consisting only of the intercept, that there is less than one chance in 10,000 that such an apparent improvement could be due only to experimental error. This strongly suggests that the first-order terms are significant, and that adding them to the model would improve it.

Looking at successive rows in the report in Fig. 5a, the same type of analysis reveals that adding the 15 two-factor interaction (2FI) terms to the six linear terms improves the model significantly (again, p-value < 0.0001), and adding the six pure quadratic terms likewise results in a substantial improvement. Note the line labeled "Cubic vs. Quadratic," though. Here, the p-value is 0.8133, meaning that the additional improvement that can be realized by adding third-order terms to this model is so small that there is over an 80% chance (81.33%) that an effect that small could be attributable to ordinary fluctuations in the experimental data. So, notwithstanding the fact that a third-order model would be of full rank and thus estimable, this analysis shows that the improvement over a quadratic model would be negligible. Likewise, a 4^{th}-order model would add no significant improvement over a cubic model.

The largest model that DX7 can analyze has 400 terms, as noted above. A 5^{th}-order model in six variables requires 462 terms, exceeding this limit. Therefore, models of 5^{th}-order and higher are excluded. Based on the analysis of 3^{rd}-order and 4^{th}-order models, however, it is unlikely that models of higher order would improve significantly on the quadratic model, even if they could be estimated.

Figure 5b is the Lack of Fit Tests report for Design Expert. For each of the models listed, the residual variance is partitioned into pure error and lack of fit components. The pure error component can only be calculated if there are replicates in the data set. In this case (Data Set 1), there were 324 pure error degrees of freedom, more than plenty to estimate the component of unexplained variance attributable to ordinary chance variations in the data. In this case, Fig. 5b shows that the pure error mean square (σ^2) is 0.30, suggesting that the standard deviation in replicated Normal Force measurements was 0.53 microV/V.

The column labeled "F Value" is the ratio of the lack of fit (LOF) mean square to 0.30, the pure error mean square. The LOF mean square is based on the difference between the predicted response and the average of all measurements made for each combination of independent variables in the test matrix. Those differences are squared and then summed to produce the total error sum of squares, from which the pure error sum of squares is subtracted to generate the LOF sum of squares. The LOF mean square is then calculated by dividing this by the number of LOF

degrees of freedom. This is just the number of unique data points in excess of the minimum number needed to fit the model.

The pure error mean square is a measure of how much error is due to the intrinsic variability of the measurement environment. The LOF mean square is a measure of how much error is due to imperfections in the way the model fits the data. The ratio of these two is the F-statistic listed in Fig. 5b for various models. A high LOF F value implies that the error is dominated by lack of fit, suggesting that the model might profit from the addition of further terms. It is unlikely that the F value will be substantially less than 1 except by chance, since we cannot expect the model to have much less variance than the data upon which it was constructed. F values near 1 suggest a good fit to the data. The corresponding p values represent the probability that an F value of the indicated magnitude could occur due simply to chance variations in the data. High p LOF p values mean that it is likely that the estimated lack of fit is due to nothing more than random fluctuations in the data, and that the model fits well. Low p values imply significant lack of fit. Note that the linear model and the two-factor interaction model both feature significant lack of fit, while the quadratic and higher-order models all fit the data quite well.

Figure 5c displays the DX7 Model Summary Statistics report. For each of the models examined, this table lists the standard deviation of the residuals, the PRESS statistic (predicted error sum of squares), and three related variations of the R-squared statistic.

To compute the PRESS statistic, the data are fitted without the first point in the test matrix. The resulting model is used to predict the first point and the difference between the measured response and that prediction is squared. This process continues for each data point and then all of the squared residuals are summed. Large PRESS statistics imply a poor fit, and reveal circumstances in which "influence points" may be driving the fit.

The R-squared statistic is computed by partitioning the total sum of squares of the data set into explained and unexplained components. The total sum of squares simply represents the sum of squared differences between each measured response and the average of all responses. The error (or unexplained) sum of squares is the sum of squared differences between each measured response and the model prediction for that point. The explained and unexplained sum of squares add to form the total sum of squares so the explained sum of squares is conveniently computed by subtracting the error sum of squares from the total. The R-squared statistic is simply the ratio of the explained sum of squares to the total.

In a perfect world devoid of unexplained variance in experimental data, all variance in the data would be explained by the model and the R-squared statistic would be exactly 1. The precision of measurements in modern aerospace calibration laboratories is in fact so great that the unexplained variance, which is responsible for experimental uncertainty, is indeed quite small, and almost all of the variance in a data set is explainable by a well-fitted model. In Fig. 5c, you can see that the R-squared statistics for all models have a value of 1 to 4 decimal places.

The Adjusted R-squared statistic displayed in the DX7 Model Summary Statistics report of Fig. 5c is competed in a similar was as the ordinary R-squared statistic, except that each sum of squares component is first divided by its corresponding number of degrees of freedom. So the "adjustment" is to ratio the explained and total mean squares (variances) rather than the sum of squares alone.

The Predicted R-squared statistics are computed using the PRESS values. The predicted error sum of squares (PRESS) is divided by the total sum of squares and the result is subtracted from 1. While the ordinary and adjusted R-squared statistics describe how well the model fits the *current* data, the predicted R-squared statistic provides some insight into how well the model will fit *new* data governed by the same relationship that has been modeled.

One unfortunate attribute of the ordinary R-squared statistic is that it continues to increase as terms are added to the model, even if those terms are insignificant. For this reason, the first author gives greater weight to the adjusted R-squared and predicted R-squared values. The adjusted R-squared tends to flatten out as insignificant terms are added to the model, and the predicted R-squared will start to decrease when too many insignificant terms are added. We prefer that the adjusted and predicted R-squared values be close to each other.

The R-squared statistics in Fig. 5c reveal that even a simple first-order model accounts for almost all of the variance in Data Set 1. This indicates that the balance is very nearly linear. However, the sequential sums of squares and the lack of fit analyses on Figs. 5a and 5b indicate that improved fits are available by adding up to second order terms, with higher order terms making a relatively small incremental contribution.

Note that in all three reports in Fig. 5, the quadratic model is underlined and labeled as "Suggested." This suggestion by the software is based on a subjective scoring system that uses a combination of selected metrics to propose an initial starting model from among those that are estimable. DX7 automatically defaults to this model, which becomes what BALFIT describes as the "Permitted Model."

To identify the Suggested model, each model is assigned two scores, as follows:

Score1 = (*M*)(*L*)(Pred R-Squared)
Score2 = (*M*)(*L*)(Adjusted R-Squared)

M is a Sequential Model Sum of Squares index and *L* is a Lack of Fit index, where

- *M* = 1 if the p-value from the sequential sum of squares report (Fig. 5a) is less than or equal to 0.05
- *M* = 0.5/(p-value) if this p-value is greater than 0.05
- *M* = 0 if model is not estimable

and

- *L* = 1 if the p-value ≥ 0.10 from the Lack of Fit report of Fig. 5b (or if Lack of Fit is not present)
- *L* = (p-value) / 0.10 if this p-value is less than 0.10

Predicted R-Squared and Adjusted R-Squared are simply the corresponding values from the Fit Summary table.

The suggested model is the one with the highest Score1, but if one model is highest on Score1 and a different model is highest on Score2, then both models will be "Suggested" and the user must choose between them. The program suggests a model with only the intercept term (the mean model) if all the predicted R-squared values are negative, or if all model scores are zero.

The user has the option to override this suggested model by selecting any of the other estimable models as a starting point. For example, in this case selecting a third-order model might permit some mixed cubic terms to enter the model that could improve the fit, notwithstanding the fact that the full cubic model provides little improvement over the full quadratic model. Truly insignificant higher order terms tend to be rejected in the stepwise regression process to follow. On the other hand, the addition of higher-order terms can simply result in a better fit to noise in the current data set, rendering the model less useful as a prediction tool. In the end, some subject matter expertise and a certain amount of prudent judgment must be brought to bear. In general, the user seeks models with these properties:

1) the highest order that is estimable;
2) no lack of fit (LOF p-value > 0.10);
3) reasonable agreement between Adjusted R-squared and Predicted R-squared (within 0.2 of each other).

B. Phase II: Model Selection

The "Suggested" model from Phase One: "Fit Summary" becomes the default model on the "Model" screen of DX7's graphical user interface. Figure 6 shows this screen for the current example, in which we are fitting the Normal Force response data from Data Set 1. The default model is the quadratic model recommended from the initial fit summary (Fig. 5), and appears as the selected option in the "Process Order" drop-down menu of Fig. 6a. The user is free to select other models from this menu to override the default.

The terms for whichever model is chosen are marked below the dropdown menus with the capital letter "M," which can be toggled on or off by the user to include or exclude individual terms. In this case, all 28 terms for a 2^{nd}-order model in six variables are included in the initial model. Terms of third order and higher (accessible by scrolling the display) are not marked for inclusion in this example, but the user can easily toggle individual terms into the model.

The user is now ready to examine subsets of this initial ("Permitted") model. Four methods are accessible from the "Selection" drop-down menu. The default is "Manual," in which the user can simply toggle individual terms in and out of the model as just described. This method is not typically invoked at this early stage, but is useful later to "fine tune" the final model.

The remaining three model selection methods on the drop-down menu are variations of a general class of methods know as "stepwise-type procedures." The three methods are known as (1) "Forward Selection," (2) "Backwards Elimination," and (3) "Stepwise Regression," which is actually a combination of the first two methods.

Forward Selection begins by assuming that only the intercept term is in the model. It then provisionally adds the one term that has the highest correlation with the response. A model F-statistic and its associated p-value are computed for this two-term model. If the p-value is below a specified threshold (indicating low probability that the added term is insignificant), the term is retained and the forward selection process continues. The next term to be provisionally entered is the one that has the highest correlation with the response after correcting for first term

entered. This is called a "partial correlation," and the associated p-value indicates the probability that this term makes an insignificant change to the explained variance of the model. If it is below the entry criterion, this term is retained. The process continues until either all terms from the initial ("Permitted") model have been included, or until the most significant remaining term does not cause enough of a change in the explained variance to satisfy the entry criterion.

Backward Elimination is similar to forward selection except that the process attempts to identify terms for the final model by working in the opposite direction. The backward elimination method begins with all terms from the Permitted Model included. The term with the weakest correlation with the response is provisionally rejected, and the impact on the explained variance of the model is assessed by an F-test. If the rejection of this term produces a significant reduction in the model's explained variance, it is retained and the process stops. Otherwise, the process continues until no terms in the model can be rejected without causing a significant reduction in the model, or until the only remaining term is the intercept.

Stepwise Regression is a combination of forward selection and backward elimination. We begin as a forward selection process and continue until the model contains the intercept and two regressors. Now apply backward elimination to the three-term model, eliminating each regressor in turn to assess the reduction in the model's explained variance via an F-test. When the backward elimination step is finished, resume the forward selection process with the whichever remaining term causes the most significant increase in the explained variance. If such a term increases the explained variance by a user-specified threshold amount, retain it and initiate the backward elimination on the new model. Continue until there are no candidate regressors significant enough to enter the model and none in the model that are so weak that they can be eliminated with no significant effect.

Proponents of stepwise regression note that the backward elimination step protects against multicollinearity. If two regressors are highly correlated, adding one of them to the model may render the first one superfluous. In such a case it is better to eliminate that term.

It is common to apply more than one stepwise procedure to the same data set. For example, the first author commonly applies backward elimination first, to give all model terms a chance to be included. He then applies stepwise regression to the surviving model terms, and finishes with one more application of backward elimination, to reject the occasional high-order term that survives the first two applications without contributing significantly to the model.

Figure 6b shows the DX7 Model Screen after the application of these stepwise procedures to the current example. This is what is called the "Recommended Model" by BALFIT. While all first-order terms were retained, several of the two-way interaction terms were rejected as insignificant, as were a couple of the pure quadratic terms. The model has been reduced from 28 terms to 17 by eliminating terms that made no significant contribution to the model's ability to predict responses, but which each carried some component of prediction uncertainty. By Eq. (13), we see that this reduced the average prediction variance to 17/28 = 61% of what it would gave been had we used the full, 28-term "Permitted Model." To achieve the same reduction in average prediction variance with the original 28-term model would require a 65% increase in data volume, with attendant increases in cycle time and direct operating costs. For the 729-point manually-loaded calibration represented by Data Set 1, this would translate into more than two additional weeks for the calibration.

The reader is advised to consult standard textbooks on regression analysis[11,12,16,28] for a more detailed description of the stepwise procedures outlined here. It is not necessary to understand the algorithms in detail to use these methods, however; one can simply choose a method and the software will execute the associated algorithm.

The stepwise procedures end with a display of the analysis of variance page in DX7. This leads to Phase Three of the variable selection and model building process.

C. Phase Three: Analysis of Variance

Figure 7a is an ordinary analysis of variance (ANOVA) table from DX7 for the current example, which corresponds to the model in Fig. 6b. This is the "Recommended Model" for Normal Force, using Data Set 1. DX7 generates an ANOVA table to display components of the total variance in a sample of data. This is useful for examining the model produced by the automated stepwise procedures described above, and provides an opportunity to inject human subject matter expertise and judgment into the model building process.

The concept exploited by an analysis of variance is that the entire data sample is characterized by variance, most of which can be explained by the math model. Some residual variance remains, which is responsible for uncertainty in the modeling result. An ANOVA partitions both the explained and the residual (or unexplained) variance into components that provide additional insights into the underlying process, which are useful for improving the model.

Figure 7a shows that the ANOVA table is organized into rows, each of which describes one component of either explained or unexplained variance, and columns, each of which provides a different quantitative descriptor of the

variance components. The first column simply labels the source of variance. The second, Sum of Squares, is divided by the third column, *df* (degrees of freedom), to produce the fourth column, Mean Square, also known as the variance component for that source. The computational details are available in standard references,[29] but the larger the mean square in column 4, the more influence that source of variance has. The variance components are presented in the next column as multiples of the unexplained variance, or residual mean square. This ratio, the F Value, is a convenient dimensionless metric of the relative importance of each source of variance.

The Model F value, at the top of the table, simply represents the ratio of the variance explained by the complete model to the unexplained variance. Small values would imply that the model has fitted mostly noise. In practical aerospace applications, however, the precision of the measurements generally guarantees a sufficient signal-to-noise ratio to produce a significant model, and the model F Value will be quite large, as in this instance.

Note that the largest component F value is the normal force load, factor A in the ANOVA table. This is consistent with expectations that the normal force output described by the model would be influenced much more heavily by the normal force load than any other source of variation in the data. (For a perfect balance, all other component F values would be zero.)

The F values reveal that for this balance, pitching moment has the greatest influence on the normal force output other than normal force itself. Even so, this variance component is five orders of magnitude less than the principal load component. Other observations available from the F values include the fact that axial force and rolling moment loads have approximately the same influence as each other, both somewhat less than the influence of pitching moment, and that by comparison, yawing moment and side force have relatively little influence on the normal force output. A significant interaction is also revealed by the high F value of the DF term, indicating that the influence that side force loads have on the normal force output depend on what the rolling moment load is, and conversely. Also note that the pure quadratic normal force load term is significant, indicating some curvature in the normal force response, but that this term is small compared to the first-order normal force loading term, implying a high degree of linearity in this transducer. Such insights can provide the balance engineer with considerable insights into the performance of instrument such as this one.

The p values in the last column of the ANOVA table express the probability that an F value as large as indicated could occur simply due to chance variations in the data sample. Small p values are associated with large F values and indicate a low probability that the apparent influence of the indicated term is due to chance. Therefore, 1-p indicates the probability that the term is real and belongs in the model.

A glance at the ANOVA table in Fig. 7a reveals that almost all the terms in the model have p values that are less than or equal to 0.0001. This implies a high probability that the true coefficient for these terms is non-zero and that we are therefore justified in retaining them in the model. The first-order yawing moment term, factor E, is a conspicuous exception, with a p value of 0.8407 and a value of F that is well below 1. This term is therefore quite unlikely to be significant, but it is retained to maintain hierarchy, an important concept discussed in Appendix A. Retaining this term renders the model "well formulated," as the linear yawing moment term is "hierarchically inferior" to the two-way interaction between yawing moment and rolling moment (the "DE" term in the ANOVA table), which is significant. The reader is referred to Appendix A for further discussion of well-formulated models and hierarchically inferior terms.

The unexplained variance is partitioned in the ANOVA table, just as the explained variance is. There are two components, "pure error" and "lack of fit." These represent the degree to which the magnitude of model residuals can be attributed to random error (imperfections in the data), and fitting error (imperfections in the model). The lack of fit F value is the ratio of the lack of fit mean square to the pure error mean square. The lack of fit p value is the probability that the degree of lack of fit estimated for this model could be due to chance variations in the data. It is quite high (0.9362) in this case, implying that it is unlikely there truly is significant lack of fit in this model. In general, the lack of fit can be regarded as insignificant when the lack of fit p value is greater than 0.1.

A number of summary statistics are presented at the bottom of the ANOVA table. Most of these were described earlier. The Mean is just the average of all the response data and the coefficient of variation (CV) is just the ratio of the standard deviation to the mean. For a nearly symmetric loading schedule, these numbers are not very meaningful as the average response is near zero, artificially inflating the coefficient of variation. The Adequate Precision number is simply the ratio of the dynamic range of the responses to the standard deviation. Large numbers imply sufficient signal to noise to fit a reasonable model. A value of 4 is considered adequate to develop a model, so the value in Fig. 7a in excess of 53,000 implies an ample signal-to-nose ratio.

Figure 7b is a standard table from DX7 that provides information for each coefficient in the present example. The coefficients are given for the model in coded units (see Appendix A). The standard error for each coefficient is listed, as are upper and lower limits for the 95% confidence interval. Note that the 95% confidence interval for the linear yawing moment term ranges from -0.11 to +0.13, a range that includes zero. This implies that we cannot say

with at least 95% confidence that this coefficient is different from zero. This is consistent with the low F and high p values in the ANOVA table that revealed this term to be insignificant, having been retained only to ensure a well-formulated model by preserving hierarchy, per recommendations in Appendix A.

The far-right column in Fig. 7b lists "variance inflation factors," a measure of multicollinearity. Multicollinearity occurs when two or more regressors are correlated to some degree. If that is the case, the model may have difficulty predicting responses for independent variable combinations other than those used to fit the model.

For the ideal case in which a regressor is perfectly orthogonal to all other regressors in the model (no colinearity), its VIF value will be 1. A standard criterion is that if VIF values are less than 10, multicollinearity is not a serious problem. A somewhat more stringent standard embraced by some researchers is that VIF values should be less than 5. By even this more stringent criterion, Fig. 7b indicates that the current model has no serious multicollinearity problems.

The ANOVA table and the table of coefficients are used to manually fine tune models developed by the stepwise procedures described above. High-order terms with low p values are candidates for deletion and can be toggled out of the model to see how this affects the standard deviation, the lack of fit p value, and the various R-squared statistics. Likewise, when an abnormally high VIF value indicates multicollinearity, higher order terms with high VIF values can be provisionally toggled out of the model to see if this relieves the multicollinearity problem without introducing significant lack of fit. In general, manual adjustments are made to the model when there is significant multicollinearity, or when insignificant terms can be deleted without violating hierarchy. With each proposed adjustment, the various quality metrics in Fig. 7 can be monitored conveniently to assess the impact.

D. Phase Four: Diagnostic Evaluation.

Except for information that may be available externally, all information on the quality of a response surface modeling experiment is contained in the residuals. For this reason, commercial software packages typically provide numerous tools for evaluating the residuals. We illustrate this in Phase Four of a typical DX7 analysis called the diagnostic evaluation phase, continuing with the Normal Force model fitted from Data Set 1 as an example. DX7 presents a total of 32 different plots of residuals for this model, and tabulates a number of them as well. A comprehensive exposition is beyond the scope of this paper, but certain key diagnostic plots of residuals are presented and will be discussed in this section.

Figure 8 presents normal probability plots of residuals for four different models fitted from Data Set 1. Normal probability paper is constructed so that a cumulative Gaussian probability distribution appears as a straight line. If the distribution of residuals falls on a straight line when plotted on such probability paper, it indicates that they are normally distributed, which suggests that they are due primarily to ordinary random error in the data, and not to some systematic imperfection in the math model. That is, a straight line implies that the model "goes through the middle" of the data, with the residuals due only to random error.

Figure 8a is the normal probability plot of residuals from the 17-term "recommended model" of Fig. 6b. The straight line suggests that the fit is adequate. Compare with Fig. 8b, which displays residuals from the 28-term full quadratic "permitted model" of Fig. 6a. This comparison reveals no loss in the quality of the fit from significantly reducing the number of model terms.

Figures 8c and 8d show how the normal plot of residuals looks when an inadequate model is fitted. Figure 8c displays residuals from a linear model featuring only the intercept and the six first-order load terms. One might have thought that such a model would fit the data reasonably well, given how small a contribution the higher-order terms make to the explained variance according to the ANOVA table. In truth, the first-order model does account for most of the variance in the data set, per the R-squared statistics of Fig. 5c. Figure 8c simply illustrates how sensitive the normal probability plot of residuals is for detecting lack of fit. Figure 8d illustrates this same point. Here, the residuals of the two-factor interaction model are plotted, which differs from the full quadratic model of Fig. 8a by only six relatively small quadratic terms. Nonetheless, there is a substantial systematic departure from a straight-line plot of the residuals, indicating lack of fit induced by the absence of the quadratic terms. Figure 8 illustrates that the normal plot of residuals can be a very sensitive indicator of lack of fit.

It is a standard practice in commercial data analysis software packages to plot residuals against predicted response levels. Such plots are expected to exhibit no functional dependence on the level of the predicted response. If this is not the case—if the residuals are proportional to the predicted response, say—then they might display a triangular shape, growing from left to right in such a plot. It is important to know this because the least squares calculations assume a uniform variance for all data points. If this assumption does not hold, the coefficient estimates will not be unbiased estimators of the true model coefficients. In such a case, a "variance stabilization transformation" of the response variables often addresses the problem, as indicated at the start of this section on

stepwise procedures. A fit of the natural log of the response often has a more nearly constant variance when the untransformed variance depends on the magnitude of the response, for example.

If the model fits the data poorly, a quadratic or higher-order dependency of the residuals will be apparent when they are plotted against predicted levels. To see why that is so, consider a math model with a poor fit. Let

$$y = f(\mathbf{x}, \mathbf{b}) \tag{14}$$

be the current (poor) model of a response, y, as a function of a vector of independent variables, $\mathbf{x}$, using a set of regression coefficients, $\mathbf{b}$.

Following Box and Draper,[12] assume that this model fits some *other* set of response data well, instead of the current data. That is, assume there is some response, $\mathbf{w}$, an unknown function of y, which this model fits better:

$$w = f(\mathbf{x}, \mathbf{b}) \tag{15}$$

Even though $\mathbf{w}$ is an unknown function of y, we can expand it over a suitably limited range as a Taylor series. If the data set that would correspond better to the current model is sufficiently close to the current data set (that is, if the current model is only marginally suboptimal), we can neglect third-order terms and higher terms in the Taylor series, as follows:

$$w = a_0 + a_1 y + a_2 y^2 \tag{16}$$

Equate (15) and (16), divide by a_1, and absorb a_0 into the intercept term of the coefficients vector, $\mathbf{b}$. Renaming the coefficients leads to a function in this form:

$$y = \alpha y^2 + f(\mathbf{x}, \mathbf{b}') \tag{17}$$

This relationship holds for all n data points, so we have

$$y_i - f(\mathbf{x}_i, \mathbf{b}') = \alpha y_i^2, \qquad i = 1,2,\ldots,n. \tag{18}$$

The term on the left can be estimated by residuals of the fitted model, and the quantity y_i on the right can be estimated from model predictions. So if the current model is not a good fit and would therefore fit another data set better than the current one, we would expect to see a characteristic quadratic dependency in the plot of residuals against model predictions, per Eq. (18).

Figure 9 shows the difference between the plot of residuals against model predictions for the recommended model (Fig. 9a), and the plot of residuals against model predictions for the factor interaction model (Fig. 9b), consisting of the full 28-term permitted model less the six pure quadratic terms. We have already established in Phase One, the Fit Summary phase, that the factor interaction model is inadequate. This was confirmed in the normal probability plots of residuals in Fig. 8. Note the characteristic quadratic shape of the residuals in Fig. 9b, confirming Eq. (18). No such quadratic dependency is apparent in Fig. 9a, suggesting that this model may be an adequate fit to the data. Third-order and higher dependencies can be seen in the plot of residuals against predicted responses if the current fit is sufficiently poor. Such plots are produced automatically by DX7 for all response models, as an additional test for model adequacy.

A plot of residuals against run number is also a staple of commercial data analysis packages. Figure 10 is an example from DX7. Run number serves as a surrogate for time. If the residuals are not a featureless function of time (that is, if they are not independent of time), it suggests that changes were occurring during the experiment that affected the response measurements.

Compare Fig. 10a, the residuals of the recommended math model developed from Data Set 1, to Fig. 10b, the residuals of the recommended math model developed from Data Set 2. Data Set 2 differs from Data Set 1 only in one respect. The experimental error simulated in Data Set 2 has a pronounced systematic component that is lacking in Data Set 1, as Fig. 1 shows. A systematic component of the unexplained variance results from bias errors that change with time. As noted earlier, these systematic errors can be caused by temperature effects, instrument drift, operator fatigue, and any number of other non-random sources of unexplained variance manifest in a time series of

measurements. The nature of the specific systematic error modeled in this study is that it causes early measurement to be biased low and later measurements to be biased high. This is reflected in the plots of residuals versus run number in Fig. 10. (There are, of course, an infinite number of variations on this theme, as the bias errors responsible for systematic components of the unexplained variance can change with time in any number of ways.)

If there are systematic errors in a data set as well as ordinary random errors, it is problematic for a response surface experiment because the true response must then be considered a function of all the known independent variables plus one additional variable, time. Under such circumstances, the response model will be just as inadequate if time is ignored as it would be if any other independent variable affecting the response is ignored. In a balance calibration experiment, systematic error means the math model depends on seven independent variables—the six load components plus time. If the response is only modeled in terms of six of these, the model will not predict responses as accurately as it would if the systematic error is taken into account. (We note in passing that it is not a simple matter of modeling the response as a function of time as well as the other independent variables, because systematic errors are transient and generally non-repeatable. The problem of systematic unexplained variance is most effectively attacked during the acquisition of the data, by invoking certain quality assurance tactics in the design of the experiment that eliminate the systematic error by converting it to another component of random error.[30-32] These tactics are key elements of the Modern Design of Experiments and are reflected in Data Sets 3 and 4.)

Unfortunately, if systematic errors are present in a calibration data set or in a data set for any other response surface modeling experiment, recovery is difficult if the experiment was not designed from the beginning to defend against this. Under such circumstances, the errors associated with successive measurements in a time series are not independent. That is, if systematic errors cause the i^{th} measurement to be too high, say, then the $(I + 1)^{st}$ measurement is also more likely to be high than low. This loss of independence violates one of the fundamental assumptions upon which regression analysis is based, bringing into question the validity of any model produced by this method. When the independent variable levels are changed systematically with time, their effects become confounded with the systematic error sources. That is, if the experiment is designed in such a way that the true change in response is due to changes in two independent variables (time and something else), but modeled as a function of only one of them, the coefficient for the one variable taken in to account will be in error. (The solution to this problem is to ensure that the independent variable levels are not changed systematically with time, but are set in random order. This is one of the key quality assurance tactics of MDOE.)

Notwithstanding the difficulty in recovering from systematic errors in an experiment that was not designed to anticipate them, it is important at least to know if they were present during the data acquisition. Otherwise it is possible to invest an unjustifiable level of confidence in the resulting math model. Plotting the residuals against run number provides valuable reassurance that there is no evidence of systematic error when that is the case (as in Fig. 10a), and provides an important warning when systematic error is present as in Fig. 10b.

Equations (14) through (18) describe a situation in which a transformation of the response variable might have resulted in a better model. Commercial data analysis packages such as DX7 provide a direct test on the data for whether such a transformation would be helpful. This method is known as the Box-Cox transformation test,[33] which examines scaled transformations involving a power transformation, y^{λ}, of the response variables, y. A model is fitted for a range of candidate exponents, typically from -3 to +3. The residuals (or the log of the residuals) are plotted against λ and the point at which this function is a minimum is a maximum likelihood estimator of the exponent that produces the best fit in a least-squares sense. Figure 11 is a representative Box-Cox transformation plot for the models examined in this study. Note that it has a very sharp minimum at $\lambda = 1$, corresponding to a recommendation of no transformation. This was true of all of the recommended models developed by DX7, which gives additional reassurance that those models are adequate.

While the Box-Cox transformation test provided no suggested transformations that would improvement in balance calibration models, it did help refine the comparisons among models developed in this study that were constructed by different analysis means, using different experiment designs, under different noise conditions. This will be illustrated presently.

This subsection has described a number of diagnostic plots of residuals that are useful in assessing model adequacy. Dozens of additional plots and tables of model adequacy indicators are provided in DX7 and other commercial data analysis packages in order to provide the opportunity to inject human expertise and judgment into the model-building process. The typical process involves a shuttling back and forth between model revision and model assessment until the response models satisfy the user. Modern commercial data analysis packages make this process quick and painless.

For most of the models developed in this study using commercial software, the initial models were adequate and the diagnostic evaluations of the residuals served simply to confirm this conclusion, with no model revisions

proposed. In some cases, however, small changes in the models improved them significantly. For example, models that initially contained highly correlated regressors could be improved by eliminating one or more of those regressors from the model. This is justified because under such circumstances, the correlated regressors contain redundant information. The resulting models were more generally transferable to other load combinations besides those that were fitted to produce the model.

As another example, occasionally the automated stepwise procedures would produce a model containing one or more terms that the ANOVA page indicated was insignificant. When these could be removed while maintaining hierarchy, they were manually eliminated. A quick glance at the subsequent diagnostic indicators would confirm in such instances that the adjustments had no unintended consequences.

A third example of circumstances in which model adjustments can make improvements is when a term is added to the recommended model based on subject-matter expertise. For example, let us say that it is known from the design of the balance that the degree of nonlinearity in one component, say "A," depends on the load applied to another component, say "B." In that case, the addition of a mixed cubic term of the form A^2B can often improve the model. Since this was a simulation experiment, no such additions were made. But this is not uncommon when data have been acquired using physical balances with known properties.

E. Summary of the Stepwise Procedures

This section has described the implementation of stepwise regression procedures in a specific data analysis package, to illustrate what is typically available in such packages and to provide a comparison with the BALFIT analysis process described in the previous section. To summarize this process, the independent variables are first specified and upper and lower limits are defined to facilitate coding of variables. Factor interaction models and all full d^{th}-order polynomial models (d from 1 to 5 in this example) are automatically examined to identify a subset of permitted models (full-rank hat matrix so nonsingular) that the test matrix will support. A weighted combination of model adequacy statistics is used to suggest an initial full model. The user can edit this or select another full permitted model as the starting point for the variable selection and model building process. Terms from the starting model are discarded or retained by the stepwise procedures, which seeks to minimize unexplained variance and multicollinearity. Hierarchy is imposed on the resulting reduced model at the user's option. Multiple metrics for fit quality are tabulated and displayed graphically. The user edits the final recommended model based on these metrics to incrementally improve the fit, to resolve multicollinearity issues, or to reflect the user's judgment and subject matter expertise.

V. Comparison Strategy

Calibration models are compared in this paper on the basis of three factors: the design of the calibration experiment, the nature of the noise environment, and the software analysis tools and methods used to generate the calibration models. The comparison strategy is described in some detail in this section, and illustrated with one specific model quality assessment metric; namely, the standard deviation of residuals between simulated measured responses and the responses predicted by various models derived from different data sets. Similar comparison analyses are reported in Section VI for other model quality metrics.

We begin this section with a description of the basic organization of the study into a two-level full factorial experiment design. This is followed by subsections describing how main effects and interactions are quantified, and how each effect is objectively classified as "significant" or "insignificant." Subsequent subsections describe interaction effects and distinguish between significance criteria that involve individual effects and combinations of effects. There is a subsection justifying a transformation of variables that under certain circumstances can improve the precision of some comparisons by increasing the signal-to-noise ratio, and there is a subsection that addresses the partitioning of explained variance components and the insights to be achieved from doing so. We end with s consideration of multicollinearity and how it is addressed in quantifying the dependencies of response models on experiment design, noise environment, and analysis method.

A. Organization as a Two-Level Full Factorial Experiment Design in Three Factors

Calibration models were developed for this study by altering three factors: the design of the calibration test matrix, the noise environment under which the data were acquired, and the software/procedures used for the analysis. There were two levels of each of these three factors.

The two levels of the DESIGN factor were "OFAT" and "MDOE." OFAT (One Factor At a Time) is a conventional calibration design represented in this case by a 729-point test matrix first introduced at Langley Research Center over 50 years ago and used with little modification ever since. MDOE (Modern Design of

Experiments) is a new calibration design based on a low-cost, high-quality research process introduced to various aerospace disciplines at Langley Research Center in the mid 1990s.

The two levels of the NOISE factor were "Random" and "Random + Systematic." Random error was modeled as a normal distribution with mean of zero and a standard deviation based on the analysis of physical replicates from a real (non-simulated) balance currently in use at Langley Research Center. An error term was selected at random from this distribution for each simulated balance response measurement. The other level of the noise factor included a component of systematic error as well as the random error, modeled as a variation in the mean of the random error distribution with time. A uniform rate was modeled that was the equivalent of just under half a standard deviation per day for the length of a typical hand-loading calibration experiment.

The two levels of the SOFTWARE factor were "DX7" and "BALFIT." DX7 is a representative commercial software package (Design Expert®, version 7) that implements standard stepwise procedures to fit general response models. BALFIT is a customized balance calibration package developed at Ames Research Center to automate the analysis of balance calibration data.

The comparisons of the factors in this study turn on examining three main effects and their interactions. We define the "main DESIGN effect" for a specified model adequacy metric as the difference in that metric between MDOE and OFAT designs, averaged over all combinations of noise environment and analysis software for the six response variables modeled. An example will clarify this below. Since there are two noise environments and two analysis software packages, there are four combinations of these variables for each of the six response variables, or 24 MDOE-OFAT differences averaged to estimate the main Design effect.

Similarly, the "Noise Effect" for a specified metric is the difference in that metric between the cases of random-error-only, and random-plus-systematic error, averaged over all combinations of test matrix design and analysis software for the six response variables. Finally, the "Software Effect" for a specified metric is the difference in that metric between DX7 and BALFIT results, averaged over all combinations of test matrix design and noise environment for the six response variables.

We are also interested in interactions among these variables. For example, the noise effect might only be significant for one experiment design and not another, in which case we would say that there is an interaction between the noise and design factors. Three such two-way interactions are possible, Noise-Design, Noise-Software, and Software-Design. Finally, there is a potential three-way interaction involving all three factors in the study. This interaction would be exist if, for example, there was a significant Noise-Design interaction for models developed by one software package, but not another.

It is convenient to assign Latin letters to the three factors as in Table 9:

Table 9. Experimental Factors.

Factor	Name	"Low"	"High"
A	Software	DX7	BALFIT
B	Design	OFAT	MDOE
C	Noise	Random	Rnd+Sys

Since each factor has only two levels, it is customary to label them as "low" and "high." The assignment of levels to these designations is entirely arbitrary and conveys no relative ranking or preference for the levels. The level assignments serve only to resolve polarity issues in the effects estimates. The "effect" for a specified model adequacy metric is defined as the change in that metric in going from the "low" level of a given factor to the "high" level. Therefore, if that effect is positive, it means the given metric *increases* with a transition from the low to the high level of the factor in question. Likewise, if the effect is negative, it means the metric *decreases* with a transition from low to high level.

With Table 9, we can use rather compact notation to describe the main effects and interactions that interest us. They are the A, B, and C main effects; the AB, AC, and BC two-way interactions; and the ABC three-way interaction. For each of the seven main effects and interactions, our objective is simply to discover whether the effect is real; that is, non-zero. For example, we will want to discover if there is any significant difference in a particular metric between DX7 and BALFIT results. If the A effect is real, we will conclude that such a difference exists. The sign of the effect will indicate which of the two factor levels is favored. In formal terms, we wish to test a *null hypothesis*, H_0, which can be expressed compactly as follows:

$$H_0: A=0$$

We will reject this hypothesis if we are able to detect some difference between the results obtained with the DX7 software and the BALFIT software. There is corresponding *alternative hypothesis*, H_1, expressed as follows:

$$H_1: A \neq 0,$$

which we will reject if we cannot detect any difference in results obtained with the two software packages. There are analogous pairs of null and alternative hypotheses for each of the seven main effects and interaction effects.

The study is structured as a two-level full factorial experiment in three factors. Table 10 shows the basic design layout of the experiment, with the standard deviation of residuals for the Normal Force model as an example comparison metric:

Table 10. Two-Level Factorial Design Layout.

ROW	SOFTWARE (A)	DESIGN (B)	NOISE (C)	NF Residual σ. μV/V
1	DX7	OFAT	Random	0.53
2	DX7	OFAT	Rnd+Sys	1.18
3	DX7	MDOE	Random	0.60
4	DX7	MDOE	Rnd+Sys	0.61
5	BALFIT	OFAT	Random	0.53
6	BALFIT	OFAT	Rnd+Sys	1.18
7	BALFIT	MDOE	Random	0.62
8	BALFIT	MDOE	Rnd+Sys	0.85

B. Estimation of Main Effects and Interactions

Two-level factorial experiment designs of the kind represented by Table 10 are quite efficient, in that they allow a large number of inferences to be made from a relatively small volume of data—eight numbers in this case. For example, note that four independent estimates of the NOISE effect are available from this sample of data. Compare rows 1 and 2. They feature the same levels of the first two factors, SOFTWARE and DESIGN, and differ only in the third factor, NOISE. We conclude therefore that any difference in the response metric—standard deviation in normal force model residuals in this instance—is attributable only to differences in the noise environment.

Recalling that a factor's main effect is defined as the response at the "high" level of that factor ("Rnd+Sys" from Table 9), less the level at the "low" level of that factor ("Random"), we have as this estimate of the main NOISE effect from rows 1 and 2: 1.18 – 0.53 = +0.65. This indicates that the addition of systematic error caused a positive change (an increase) in residual standard deviation of about 0.65 μV/V in this example. Note, however, that this result only applies to one combination of factors A and B—OFAT designs analyzed with the DX7 software. Rows 3&4, 5&6, and 7&8 likewise provide independent estimates of the NOISE effect for normal force residual standard deviation. These three estimates plus the first one cover all four combinations of the two levels of SOFTWARE and DESIGN, as summarized in Table 11. The main NOISE effect is defined as the average of these estimates, which in this example is +0.39.

We conclude that for the Normal Force models, adding systematic error of the type modeled in this study to the random error increases the standard deviation in model residuals by 0.39 μV/V averaged over all combinations of the two experiment designs and two software packages. We say that such an estimate features a "wide inductive basis" because it spans all possible combinations of the SOFTWARE and DESIGN factors, rather than one "representative" combination.

Table 11. NOISE Effects for Normal Force Residual Standard Deviation, μV/V.

ROWS (Table 10)	SOFTWARE (A)	DESIGN (B)	Noise Effect
1&2	DX7	OFAT	+0.65
3&4	DX7	MDOE	+0.01
5&6	BALFIT	OFAT	+0.65
7&8	BALFIT	MDOE	+0.23

Note that there is some variability in the NOISE effects presented in Table 11. For example, for both DX7 and BALFIT, the addition of systematic error apparently inflates the Normal Force residual standard deviation more when the model is developed from an OFAT experiment design than when it is developed from an MDOE design (0.65 vs. 0.01, and 0.65 vs. 0.23, respectively). This reflects the quality assurance tactics embedded within the MDOE design to defend against systematic errors. It also illustrates that while the average NOISE effect is 0.39 μV/V, the magnitude of the effect appears in this example to depend on the level of one or both of the other two factors. We say in such circumstances that an *interaction* exists.

To quantify the interaction between NOISE and DESIGN, we subtract the average noise effect for the low level of DESIGN (OFAT) from the average noise effect for the high level of DESIGN (MDOE), and normalize to facilitate a direct comparison with the main effects. We denote this interaction by "BC."

The same calculations can be performed for the DESIGN and SOFTWARE main effects and their interactions as have been computed for the NOISE factor. For example, rows 1&3, 2&4, 5&7, and 6&8 all provide independent estimates of the DESIGN effect and rows 1&5, 2&6, 3&7, and 4&8 do likewise for the SOFTWARE main effect.

Residual standard deviations for all six balance response components were computed for the eight factor combinations illustrated in Tables 10—a total of 48 estimates of residual standard deviation. These data were "blocked" by response, a technique that increases the number of degrees of freedom available to assess the three main factor effects and their interactions, and provides results that span all six response components. Table 12 summarizes the results, displaying all main effects and interactions blocked to include all balance output responses.

Table 12. Residual Standard Deviation Main Effects and Interactions across All Response Components, μV/V.

Effect	Value, μV/V
A: SOFTWARE	0.03
B: DESIGN	-0.41
C: NOISE	0.50
AB	-0.01
AC	0.04
BC	-0.42
ABC	0.01

C. Objective Identification of Significant and Insignificant Effects

Recall that our objective is to infer whether effects such as those in Table 12 are real or not, by which we mean "non-zero" (whether positive or negative). Effects that are of sufficient magnitude to be distinguishable from zero with a prescribed level of confidence are said to be "significant." For this study, "significant" effects are those that can be distinguished from zero with at least 99% confidence. That is, we will accept up to a 1% (0.01) probability of an incorrect inference (due, say, to experimental error) if we infer that a given effect is real. This inference error tolerance level is commonly referred to as the "significance" of the study. (Somewhat paradoxically, the smaller the "significance" by this definition, the greater the significance of the conclusions!)

In formal; terms, we wish to decide whether to reject the *null* hypothesis for each of the effects (concluding that the effect is real), or to reject the *alternative* hypothesis (concluding that there is no significant effect). Because of experimental error, we cannot simply reject the null hypothesis for every effect that is non-zero. Even if there truly is no effect, ordinary experimental error will cause all of the effects estimates to be non-zero except by rare coincidence. Some of the effects in Table 12 are clearly greater than others. The NOISE and DESIGN main effects, along with the interaction between them, seem to dominate the other effects, but it is not yet clear whether we are justified in describing these effects as significant.

One way to make objective inferences as to whether the null hypothesis or its alternative should be rejected is to rigorously compute the uncertainty in estimating each effect, and from this determine if the effect is large enough to unambiguously distinguish it from zero. If so, we would reject the null hypothesis and if not, we would reject the alternative hypothesis. However, there is a graphical method that is less tedious and more convenient. It utilizes the normal probability plot introduced in section IV-D and illustrated in Fig. 8. Recall that normally distributed points will fall on a straight line in such a plot and points that are not normally distributed will lie off of such a line. We exploit this property to distinguish between effects that differ systematically from zero, and those that own their departure from zero merely to random variations associated with experimental error. We know by the Central Limit

Theorem that ordinary experimental error is normally distributed, so we expect the latter points to fall on a straight line on normal probability paper while the former points lie off the line. Figure 12 illustrates this concept for the data in Table 7.

Figure 12 confirms our earlier speculation from examining Table 12 that the B and C main effects are significant as well as the BC two-way interaction between them. All three effects are unambiguously off the line. The triangles in this plot represent error from replicates, generated when the analysis was extended over all six response variables.

Effects to the right of the line are positive and those to the left are negative. Recall that an "effect" is defined as a change in some response (the calibration residual standard deviation in this case) due to a transition from the "low" level of some factor to the "high" level, per Table 9. The fact that the NOISE effect, C, is to the right of the line and is therefore positive, coupled with the fact that "Random" error was defined as the "low" level of the noise factor while random-plus-systematic error was defined as the "high" level, means that adding systematic error to the noise will cause the calibration residual standard deviations to increase. While this is hardly an unexpected result, it does confirm our intuitive expectations and serves as a simple illustration of the interpretation of normal probability plots.

The DESIGN effect, B, is to the left of the line in Fig. 12 and is therefore negative. From the definitions of low and high levels for this factor in Table 9, we conclude that a change in the design of the experiment from OFAT to MDOE would reduce the calibration residual standard deviations.

The interaction effect, BC, is also negative. Note that while a main effect represents a change in a *response*, a two-way interaction represents a change in an *effect*. The BC interaction represents the change in the C main effect when the B factor changes from "low" to "high." In this example, the negative BC interaction means that the noise effect, C, gets smaller as the experiment design factor, B, transitions from OFAT to MDOE. That is, adding systematic error increases the residual standard deviation for both OFAT and MDOE designs, but less so for an MDOE experiment design than for an OFAT experiment design. The interpretation of two-way interactions is entirely symmetric, so that we could just as easily say that a negative BC implies that the DESIGN effect gets larger as we transition from the low level to the high level of the NOISE factor. We say "larger" because the DESIGN effect is already negative—going from OFAT to MDOE *reduces* the residual standard deviation. For BC to have a negative sign, this means that the reduction in residual standard deviation must be greater when there is systematic error than when there is only random error. The next subsection discusses interaction effects in more detail.

Normal probability plots such as Fig. 12 are often as revealing for the effects they do *not* show, as for the significant effects they identify. Note, for example, that there is no significant A effect (SOFTWARE factor) in Fig. 12. This means that, across all six response variables, for both experiment designs, and in both noise environments, no significant difference could be detected between the calibration residuals of data sets analyzed with BALFIT and with a commercial data analysis package, Design Expert. Also, there are no significant interaction effects involving factor A. That is, the NOISE effect, the DESIGN effect, and their two-way interaction, are all estimated to be the same whether the calibration residuals are based on models developed by BALFIT or by Design Expert.

D. Elucidation of Significant Interaction Effects

Figure 13, an interaction graph, is a standard data structure used in the analysis of two-level factorial experiments such as this one. It clearly illustrates that the addition of systematic error has a much more serious effect on data acquired in an OFAT experiment than on data acquired using MDOE. The error bars on each data point in this figure represent 95% Least Significant Differences (LSD). If the LSD bars for two points overlap, we are unable to distinguish a difference between them with at least 95% confidence. Clearly there is substantial overlap of the MDOE and OFAT LSD intervals for the case of random-only experimental error. This suggests that when there is systematic error, we cannot resolve a difference between the residual standard deviation of a data set acquired with an OFAT experiment design and a data set acquired with an MDOE experiment design. There are two points to make about this. First, since much of the raison d'etre of MDOE experiment design is to defend against systematic error, it is not surprising that its greatest benefits are derived under conditions when systematic error is present. Secondly, the OFAT design consumed the resources required to obtain 729 data points, while the MDOE design required only 64 points—an order of magnitude difference, with attendant increases in direct operating cost and cycle time, and with no resolvable increase in quality. This suggests that the OFAT design is substantially more wasteful of resources than the MDOE design.

E. Pareto Charts and the Bonferroni Limit

A normal probability plot such as Fig. 12 is very convenient for quickly identifying which effects are significant and which are not. The distance each significant point is away from the straight line in such a normal probability plot reveals some information about the relative magnitude of each effect, but these relative magnitudes are

displayed much more clearly in a Pareto chart as in Fig. 14. A Pareto chart is simply a bar chart that rank orders each effect by its magnitude, expressed as a multiple of the standard deviation of the error in estimating it (the t-value).

The Pareto chart also reveals how confident we can be that each effect is non-zero. Those effects that rise above the lower horizontal line on the Pareto chart—labeled the t-Value limit—are of sufficient magnitude that we can still distinguish the effect from zero with at least 95% confidence, notwithstanding the experimental error in the data and the attendant uncertainty it causes.

There is one potential difficulty in using confidence limits as a criterion for identifying individual significant effects. Imagine for a moment that the three largest effects in Fig. 14 are significant, but just at the 95% confidence level. That is, suppose that each has a 95% probability of being real (non zero). The probability that all three effects are real is then $0.95 \times 0.95 \times 0.95 = 0.86$, well below our 95% criterion. To have a probability of 95% that *all three* effects are real requires an average individual probability of 98.3%. This may seem like a small difference from 95%, but it amounts to odds of about 58:1 vs. 19:1. The introduction of a joint-probability criterion therefore provides a much more stringent condition for interpreting effects as real.

Effects that are sufficiently large that there is a specified probability that they are *all* real are said to exceed the Bonferroni limit. For the Pareto chart in Fig. 14, the Bonferroni limit corresponds to a 95% confidence level. It is clear from this chart that all three of the effects identified on the normal probability plot of Fig. 12 as significant are large enough to exceed the Bonferroni limit, and we therefore infer with at least 95% confidence that all three are real (non-zero) effects.

The Bonferroni criterion is clearly more stringent than the t-Value criterion. While both criteria are commonly used, in this study we adopt the Bonferroni criterion, declaring only those effects to be real that have a joint probability of 95% or greater that they are all non-zero.

F. The Utility of Variable Transformations

We must digress briefly to discuss an important technical detail. The normal probability plot and Pareto chart are constructed under the assumption that experimental errors follow a Gaussian distribution, with a standard deviation that is constant for each data point analyzed. This is generally the case when the data consists of ordinary measured values. But in this case, the data consists of *computed* values—standard deviations in the regression residuals. The standard deviation follows an asymmetric probability density function with a long positive tail. The variance in the estimate is not constant when we are estimating standard deviations. Rather, it is proportional to the magnitude of the estimated standard deviation.

This common situation requires a "variance stabilization transformation" to transform the variables into something with a constant variance. The logarithmic transformation often fits the bill nicely. Assume that we analyze the *logarithm* of the standard deviations rather than the standard deviations themselves, simply for the purpose of identifying significant effects using probability plots and Pareto charts. That is, let s_i be the standard deviation of regression residuals corresponding to a math model developed for the i^{th} combination of NOISE, DESIGN, and SOFTWARE factors, and let $y_i = \log(s_i)$ be its logarithm. The variance in the distribution of s is proportional to s:

$$\sigma_s^2 = ks \tag{19}$$

We use ordinary error propagation[34] to compute the variance in the distribution of *logarithms* of s_i:

$$\sigma_{y(s)}^2 = \left[\partial y(s) \Big/ \partial s \right]^2 \sigma_s^2 = \left[\partial \log(s) \Big/ \partial s \right] \sigma_s^2 = \left(1 \Big/ s \right)^2 (ks)^2 = k^2, \text{ a constant} \tag{20}$$

So from Eqs. (19) and (20) we see that while the variance of the distribution of standard deviations is *not* constant, the variance in the distribution of *logarithms* of the standard deviation *is* constant. We would expect much more sensitive estimates of the significance of various effects using this transformation, since we no longer have to cope with the long tail of the distribution of standard deviations and its tendency to produce large outliers.

Fortunately, it is not necessary to make judgments about the efficacy of applying a transformation to each individual response variable of interest. We simply apply the Box-Cox transformation test discussed earlier and illustrated in Fig. 11. Recall that the Box-Cox transformation test objectively determines if a power-law transformation will produce better results. That is, this test indicates whether a better result can be achieved with y^{λ} rather than y, with the optimum λ indicated as the minimum of the Box-Cox transformation plot shown in Fig. 11.

The Box-Cox power transformation plot in Fig. 15 applies to the standard deviation effects data discussed in this section, and suggests that a better result can be achieved with a power transformation in which the exponent is zero. It turns out that in the limit as lambda approaches zero, y^{λ} approaches $\log(y)$. That is, for small lambda, y^{λ} plots against log(y) very nearly as a straight line.[12] This confirms conclusions based on Eq. (20), that a logarithmic transformation is expected to have better variance properties and will therefore be a more sensitive indicator of significant effects.

For this study, the Box-Cox power transformation test was applied routinely to the data for each model adequacy metric. Transformations were applied when they could improve the sensitivity of the effects tests. Figure 16 illustrates how the logarithmic transformation improved the sensitivity of the normal probability plot for identifying significant residual standard deviation effects. Note that the straight line is much less slanted, meaning that an effect of a given magnitude is farther from the straight line and therefore more easily resolvable.

In this instance, the transformation did not yield any new insights; B (DESIGN), C (NOISE), and BC are still identified as the only significant effects for residual standard deviation. But there are circumstances in which an effect would stand unambiguously off the straight line in the normal probability plot of *transformed* effects, when it would not before the transformation.

The effect of the logarithmic transformation is also seen in Fig. 17, comparing Pareto charts before and after transformation. Note from the scale of the *y*-axis that the t-values are over twice as great after transformation, indicating a substantial improvement in signal to noise ratio for assessing the significance of the effects. Note also that the four insignificant effects—A, AB, AC, and ABC—while all still comfortably below either significance criterion (t-value limit or Bonferroni limit), are nonetheless closer to those lines than in the untransformed Pareto chart. This is because the transformation has improved the signal-to-noise ratio to the point that these effects can almost be resolved in the noise, but not quite. So we retain our original inference, which is that we can only say with 95% confidence that the B, C, and BC effects are real. We reject the null hypothesis for those three effects. We reject the alternative hypothesis for the A, AB, AC, and ABC effects.

G. Partitioning of the Explained Sum of Squares

The total sum of squares for any ensemble of data can be computed by subtracting each data point from a specified reference, squaring the difference, and adding all the squared values. The reference most commonly used is the average of the data. The sum of squares is an indication of the amount of variability that exists among the different data points. In a good experiment, most of the variability will have been caused by changes imposed by altering the levels of the factors under study. We refer to this as "explained variance." After accounting for all the explained variance, however, there always remains a residual variance that is unexplained. It is because of this unexplained variance that there must necessarily be some uncertainty associated with any experimental result.

While the details are beyond the scope of the present paper, it is possible to partition the total sum of squares into explained and unexplained components. It is possible to further partition the explained sum of squares into components attributable to individual main effects and interactions among the factors. For the current example in which we are analyzing how the three factors of this study impact the standard deviation of calibration residuals, the component sums of squares are represented in the pie chart of Fig. 18. This figure reveals that in the case of residual standard deviation, the three significant effects involving the NOISE and DESIGN factors and their interaction, account for 99.6% of the explained sum of squares. Only 0.4% can be attributed to the A (SOFTWARE) factor and its interactions with noise and design. For the case of residual standard deviation, it appears that the choice of experiment design has about the same effect in this study as the noise environment.

Figure 19 summarizes the findings for residual standard deviation as a response surface modeling adequacy metric. The normal probability plot identifies significant and insignificant factor effects, the Pareto chart illustrates their relative magnitude and indicates which effects are jointly significant (above Bonferroni limit), the interaction graph provides insights into the relationship between noise and experiment design, and the pie chart shows what fraction of the overall explainable variability in the data can be attributed to the various factors and factor interactions. Composite figures such as Fig. 19 are displayed in the next major section for a variety of math model quality metrics, to show how they are influenced by the choice of experiment design, noise environment, and analysis software.

H. Multicollinearity

It is not uncommon for a regressor in a math model to exhibit some degree of correlation with one or more other regressors. When they exhibit near-linear dependencies a condition known as multicollinearity is said to exist. The consequences of multicollinearity and some methods for dealing with it are discussed in Appendix B. A particular source of multicollinearity due to a physical constraint among balance calibration loads comes into play when a

Single Vector System is used to apply the loads. Single Vector Systems are further described in Appendix B as is the constraint they introduce, which is responsible for substantial multicollinearity in models that feature all three components of the dot product between the total force load vector and the total moment load vector; namely, the (NF)(YM), (AF)(RM), and (SF)(PM) terms.

Since SVS loading was developed at Langley Research Center to accommodate MDOE calibration designs, each point in the MDOE loading design simulated in this study features the dot product constraint described in Appendix B. This constraint does not come into play unless all three of the two-way interaction terms are significant and thus retained in the model; if any of them is excluded from the model, the multicollinearity is relieved. Analysts at Langley Research Center who routinely use DX7 to analyze data generated with SVS load schedules exploit this to produce models that do not suffer from multicollinearity. However, SVS loading is not commonly practiced at Ames Research Center and the BALFIT software system developed there does not currently include a test for multicollinearity.

Unfortunately, of the 24 models developed in this study by BALFIT, four of them did feature significant terms for all three components of the total force-moment dot product and therefore displayed multicollinearity. This introduces a lack of balance in the full-factorial design of this study which would complicate the analysis if not taken into account. In order not to skew the model comparisons with multicollinearity effects that are present in some models and not in others, the analyses reported in the next section are limited to responses in which the dot product constraint did not come into play. This is the case for most of the data, and there was still ample precision available after deleting the multicollinear cases to resolve subtle effects in the three main factors of the study.

Extensions to BALFIT are currently being tested which will enable it to detect multicollinearity and automatically correct for the presence of collinear terms in the recommended model. This will enable BALFIT to analyze calibration data acquired with SVS loading systems, as well as data from the conventional calibration experiments for which it was originally designed.

VI. Comparison of Analysis Results

The analyses described in the previous section and summarized in Fig. 19 for the standard deviation of calibration residuals was applied to numerous other model quality metrics as well. Each metric is described in this section. We make seven inferences for each response metric by rejecting either the null hypothesis or the alternative hypothesis associated with the three main effects of this study, the three two-way interactions, and the one three-way interaction. These hypothesis pairs are summarized here for convenience:

H_{01}: A=0 No difference in results obtained by DX7 and BALFIT
H_{11}: A≠0 DX7 and BALFIT produce different results

H_{02}: B=0 No difference in results obtained by with MDOE and OFAT experiment designs
H_{12}: B≠0 MDOE and OFAT experiment designs produce different results

H_{03}: C=0 No difference in results when systematic error is added to the random error
H_{13}: C≠0 Results are different when systematic error is present than when there is only random error

H_{04}: AB=0 Differences in results obtained by MDOE and OFAT are independent of analysis software
H_{14}: AB≠0 Differences between MDOE and OFAT results depend on analysis software

H_{05}: AC=0 Differences attributable to noise environment are independent of analysis software
H_{15}: AC≠0 Différences attributable to noise environnent depend on analysis software

H_{06}: BC=0 Differences in results obtained by MDOE and OFAT are independent of noise environment
H_{16}: BC≠0 Differences between MDOE and OFAT results depend on noise environment

H_{07}: ABC=0 The level of any two-way interaction is independent of the level of the third factor
H_{17}: ABC≠0 The level of any two-way interaction depends on the level of the third factor

The three-way interaction hypotheses ask, for example, if the dependence of the DESIGN effect on NOISE environment is computed to be different when one SOFWARE system is used for the analysis than the other.

For each model quality metric examined in this study, one hypothesis is rejected for each hypothesis pair. Those results are summarized in this section. For each metric we provide a description, discuss its importance, present observations from the data, and draw conclusions from those observations. All conclusions are reported with at least 95% confidence in the stringent Bonferroni-Limit sense.

A. Standard Deviation of Calibration Model Residuals

1. Description and Computation

Square root of residual variance, or standard deviation of unexplained variance. Computed as the root-mean-square of differences between predicted and simulated measured responses for all points used to fit the regression model.

2. Observations

Observations for this model quality metric were reported in the last section as a detailed example illustrating the comparison methodology of this study. These observations are drawn from Fig. 19:

A1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit
A2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
A3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
A4. Adding the systematic component of unexplained variance increased the standard deviation of calibration model residuals by a statistically significant amount when the data were acquired with the OFAT experiment design, but there was no significant increase when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the standard deviation of model residuals for the models obtained by DX7 and BALFIT. This conclusion is supported by observation A3.
- We reject H_{02}: B=0, concluding that OFAT experiment designs lead to greater standard deviations of model residuals than MDOE experiment designs. This conclusion is supported by observations A1 and A2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater standard deviations of model residuals than for the case of the random component alone. This conclusion is supported by observations A1 and A2.
- We reject H_{14}: AB≠0, concluding that differences in standard deviations of model residuals obtained with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation A3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the standard deviation of calibration model residuals that are independent of whether BALFIT or DX7 was used to generate the models This conclusion is supported by observation A3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on the standard deviation of calibration model residuals depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations A1 and A4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for the standard deviation of calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE is computed to be the same whether BALFIT or DX7 is used for the analysis. This conclusion is supported by observation A3.

4. Discussion

Models developed by BALFIT and DX7 result in calibration model residual standard deviations that are essentially the same. However, for either software package, calibration quality as assessed by this metric is impacted by the addition of systematic error for OFAT experiment designs. No resolvable difference was detected between

the calibration model residual standard deviations of MDOE models generated in the presence of random noise only, and the random-plus-systematic noise illustrated in Fig. 1. This is attributed to MDOE quality assurance tactics incorporated into the test matrix design; specifically, randomization and blocking intended explicitly to defend against systematic error.

B. Maximum calibration model residual

1. Description and Computation

Largest absolute difference between predicted and simulated measured response for all points used to fit the math model. Model predictions at each point used to fit the model were subtracted from the simulated measurement at that point and the largest absolute difference was recorded for each model examined.

2. Observations

Observations for this model quality metric are drawn from Figs. 20 and 21.

B1. The A (SOFTWARE), B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit, however the SOFTWARE effect is barely resolvable and its Least Significant Difference bars slightly overlap in Fig. 21.
B2. The A (SOFTWARE) and C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
B3. None of the interaction effects involving A (SOFTWARE) is significant (AB, AC, or ABC).
B4. Adding the systematic component of unexplained variance increased the maximum calibration model residuals by a statistically significant amount when the data were acquired with the OFAT experiment design, but there was no significant increase when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in maximum model residuals for the models obtained by DX7 and BALFIT. This conclusion is supported by observation B1.
- We reject H_{02}: B=0, concluding that OFAT experiment designs lead to greater maximum model residuals than MDOE experiment designs. This conclusion is supported by observations B1 and B2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater maximum model residuals than for the case of the random component alone. This conclusion is supported by observations B1 and B2.
- We reject H_{14}: AB≠0, concluding that differences in maximum model residuals obtained with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation B3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in maximum calibration model residuals that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation B3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on maximum calibration model residuals depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations B1 and B4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for maximum calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE is computed to be the same whether BALFIT or DX7 is used for the analysis. This conclusion is supported by observation B3.

4. Discussion

Models developed by BALFIT were observed to generate slightly larger maximum calibration model residuals than models developed by DX7, although this difference was just barely resolvable in the data. See Fig. 20. The LSD bars for BALFIT and DX7 touch and slightly overlap in Fig. 21 so we are unable to unambiguously reject the

H_{01} null hypothesis. We conclude, therefore, that there is no practical difference between the maximum calibration model residuals for models developed under BALFIT and DX7.

As with the standard deviation of calibration residuals, for either software package the addition of systematic error impacted the OFAT experiment designs but not the MDOE designs. Again, this is attributed to MDOE quality assurance tactics incorporated into the test matrix design for the express purpose of defending against systematic error and the loss of measurement independence that that occurs when it is not taken into account.

C. Standard Deviation of Confirmation Point Residuals

1. Description and Computation

Computed as the root-mean-square of differences between predicted and simulated measured responses for 25 confirmation points used to test the regression model. These points were not used to fit the model.

2. Observations

Observations for this model quality metric were drawn from Fig. 22.

C1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit
C2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
C3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
C4. Adding the systematic component of unexplained variance increased the standard deviation of confirmation point residuals by a statistically significant amount when the data were acquired with the OFAT experiment design, but there was no significant increase when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the standard deviation of confirmation point residuals for the models obtained by DX7 and BALFIT. This conclusion is supported by observation C3.
- We reject H_{02}: B=0, concluding that OFAT experiment designs lead to greater standard deviations of confirmation point residuals than MDOE experiment designs. This conclusion is supported by observations C1 and C2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater standard deviations of confirmation point residuals than for the case of the random component alone. This conclusion is supported by observations C1 and C2.
- We reject H_{14}: AB≠0, concluding that differences in standard deviations of confirmation point residuals obtained with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation C3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the standard deviation of confirmation point residuals that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation C3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on the standard deviation of confirmation point residuals depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations C1 and C4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for the standard deviation of confirmation point residuals. This means that the significant interaction observed between DESIGN and NOISE is computed to be the same whether BALFIT or DX7 is used for the analysis. This conclusion is supported by observation C3.

4. Discussion

Models developed by BALFIT and DX7 result in confirmation point residual standard deviations that are essentially the same. As with other metric examined in this study, for either software package the calibration quality as assessed by this metric is impacted by the addition of systematic error for OFAT experiment designs. Again, no resolvable difference was detected between the confirmation point residual standard deviations of MDOE models generated in the presence of random noise only, and the random-plus-systematic noise illustrated in Fig. 1.

D. Maximum Confirmation Point Residual

1. Description and Computation

Largest absolute difference between predicted and simulated measured response for 25 confirmation points used to test the regression model. These points were not used to fit the model.

2. Observations

Observations for this model quality metric are drawn from Fig. 23.

D1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
D2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
D3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
D4. Adding the systematic component of unexplained variance increased the maximum confirmation point residuals by a statistically significant amount when the data were acquired with the OFAT experiment design, but there was no significant increase when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in maximum confirmation point residuals for the models obtained by DX7 and BALFIT. This conclusion is supported by observation D3.
- We reject H_{02}: B=0, concluding that OFAT experiment designs lead to greater maximum confirmation point residuals than MDOE experiment designs. This conclusion is supported by observations D1 and D2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater maximum confirmation point residuals than for the case of the random component alone. This conclusion is supported by observations D1 and D2.
- We reject H_{14}: AB≠0, concluding that differences in maximum confirmation point residuals obtained with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation D3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in maximum confirmation point residuals that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation D3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on maximum confirmation point residuals depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations D1 and D4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for maximum calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE for maximum confirmation point residuals is computed to be the same whether BALFIT or DX7 is used for the analysis. This conclusion is supported by observation D3.

4. Discussion

Models developed by BALFIT and DX7 result in maximum confirmation point residuals that are essentially the same. As with other metrics examined in this study, for either software package the calibration quality as assessed by this metric is impacted by the addition of systematic error for OFAT experiment designs. Again, no resolvable

difference was detected between the confirmation point residual standard deviations of MDOE models generated in the presence of random noise only, and the random-plus-systematic noise illustrated in Fig. 1.

E. Number of Successful Confirmations

1. Description and Computation

Twenty-five confirmation points were acquired at randomly-selected load combinations within the load range of the balance to test each model. These points were not used in any of the regression analyses to create the models. A point was considered successfully confirmed if it fell within the 95% prediction interval associated with the model being tested.

A Critical Binomial Analysis was performed to determine if the number of successful confirmations was large enough to certify the model as an adequate predictor of responses at other load combinations besides those used to fit the data. We cannot require 25 successes out of 25 attempts as the criterion for certifying the model because there is uncertainty in the model prediction and also uncertainty in the confirmation points; individual confirmation point successes are evaluated in terms of 95% prediction intervals, not 100% prediction intervals. The Critical Binomial Number is a tabulated statistic that describes the minimum number of successes one is entitled to expect a given percent of the time when there is a specified number of trials for which the probability of success for each trial is known. One can use the CRITBIMOM function in Excel to compute this number.

For the case of 25 trials in which the probability of success in each trial is assumed to be 95%, the critical binomial number is 21, with a significance of 0.01. That is, if the true response actually does lie within the 95% prediction interval limits 95% of the time, then in a set of 25 independent trials we would expect 21 or more successes to occur 99% of the time that such a 25-point test is conducted.

2. Observations

Observations for this model quality metric are drawn from Fig. 24:

E1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
E2. The C (NOISE) effect is negative while the B (DESIGN) and BC effects are positive.
E3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
E4. Adding the systematic component of unexplained variance decreased the number of successful confirmations when the data were acquired with the OFAT experiment design, but there was no significant decrease in the number of successful confirmations when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: $A \neq 0$, concluding that there is no significant difference in the number of successful confirmations for models obtained by DX7 and BALFIT. This conclusion is supported by observation E3.
- We reject H_{02}: $B=0$, concluding that MDOE experiment designs lead to more successful confirmations than OFAT experiment designs. This conclusion is supported by observations E1 and E2.
- We reject H_{03}: $C=0$, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to fewer successful confirmations than for the case of the random component alone. This conclusion is supported by observations E1 and E2.
- We reject H_{14}: $AB \neq 0$, concluding that differences in the number of successful confirmations obtainable with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation E3.
- We reject H_{15}: $AC \neq 0$, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the number of successful confirmations that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation E3.
- We reject H_{06}: $BC=0$, concluding that the addition of the systematic component of unexplained variance had a different effect on the number of successful confirmations that are achieved, depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to

the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations E1 and E4.

- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for maximum calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE for the number of successful confirmations is computed to be the same whether BALFIT or DX7 is used for the analysis. This conclusion is supported by observation E3.

4. Discussion

Models developed by BALFIT and DX7 are equally likely to be confirmed by using them to predict responses for loads that are not part of the calibration load schedule. As with other metrics examined in this study, for either software package the calibration quality as assessed by this metric is impacted by the addition of systematic error for OFAT experiment designs. For random noise only, OFAT and MDOE models were equally likely to be confirmed. No resolvable difference was detected between the number of successful confirmations of MDOE models generated in the presence of random noise only versus random-plus-systematic noise, but the introduction of systematic error dramatically reduced the probability of successfully confirming models developed with an OFAT design.

F. Lack-of-Fit F-Statistic

1. Description and Computation

The residual or unexplained variance is partitioned into a component attributable to ordinary chance variations in the data and the remainder, attributed to model imperfections. The ratio of the latter to the former is the lack-of-fit F-statistic. We do not expect this number to fall below one except for chance variations in the data, since the model is built from the data and cannot have less variance. But smaller LOF F values generally indicate a better fit to the data.

2. Observations

Observations for this model quality metric are drawn from Fig. 25.

F1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
F2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
F3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
F4. Adding the systematic component of unexplained variance increased the lack-of-fit F-statistic when the data were acquired with the OFAT experiment design, but not when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the lack of fit F-statistic for models obtained by DX7 and BALFIT. This conclusion is supported by observation F3.
- We reject H_{02}: B=0, concluding that MDOE experiment designs lead to lower lack of fit F-statistics than OFAT experiment designs. This conclusion is supported by observations F1 and F2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater lack of fit than when there is only random noise. This conclusion is supported by observations F1 and F2.
- We reject H_{14}: AB≠0, concluding that differences in lack of fit obtainable with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation F3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in estimates of the lack-of-fit F statistic that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation F3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on model lack of fit, depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the

quality of models developed with the OFAT design. This conclusion is supported by observations F1 and F4.

- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for maximum calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE for the lack-of-fit F statistic is the same whether the models are developed with BALFIT or DX7. This conclusion is supported by observation F3.

4. Discussion

Models developed by BALFIT and DX7 appear to be characterized by the same degree of lack of fit. As with other metrics examined in this study, for either software package the lack-of-fit F statistic is adversely impacted by the addition of systematic error for OFAT experiment designs. For random noise only, OFAT and MDOE models had indistinguishable lack-of-fit F statistics. No resolvable difference was detected in this statistic for MDOE models generated in the presence of random noise only versus random-plus-systematic noise, but the introduction of systematic error dramatically increased the estimated lack of fit for models developed with an OFAT design.

G. Prediction Uncertainty

1. Description and Computation

The average 95% confidence interval half-width for prediction uncertainty estimates the precision with which the model can predict responses. Based on Eq. (13), and assuming a normal distribution with sufficient degrees of freedom in variance estimates that the 95% confidence interval half width is 2σ, it is computed as follows:

$$2\sqrt{\frac{p}{n}}\sigma \tag{21}$$

where σ is the standard deviation, p is the number of terms in the math model including the intercept, and n is the number of points used to fit the model. The true response is assumed to lie within an interval centered on the predicted value and twice as wide as Eq. (21).

2. Observations

Observations for this model quality metric are drawn from Fig. 26.

G1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
G2. The B (DESIGN) and C (NOISE) effect are positive while the BC effect is negative.
G3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
G4. Adding the systematic component of unexplained variance increased prediction uncertainty when the data were acquired with the OFAT experiment design, but not when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the prediction uncertainty for models obtained by DX7 and BALFIT. This conclusion is supported by observation G3.
- We reject H_{02}: B=0, concluding that the OFAT experiment design lead to greater precision in prediction than the MDOE experiment design. This conclusion is supported by observations G1 and G2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to greater prediction uncertainty than when there is only random noise. This conclusion is supported by observations G1 and G2.
- We reject H_{14}: AB≠0, concluding that differences in prediction uncertainty obtainable with MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation G3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in prediction uncertainty that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation G3.

- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on prediction uncertainty, depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations G1 and G4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for maximum calibration model residuals. This means that the significant interaction observed between DESIGN and NOISE for prediction uncertainty is the same whether the models are developed with BALFIT or DX7. This conclusion is supported by observation G3.

4. Discussion

Models developed by BALFIT and DX7 appear to provide equally precise response predictions. As with other metrics examined in this study, for either software package prediction uncertainty is adversely impacted by the addition of systematic error for OFAT experiment designs. The prediction precision of models developed from MDOE designs are only negligibly impacted by the addition of systematic error, while the prediction precision of OFAT models is degraded significantly. However, in either noise environment the prediction precision of the OFAT models was greater than that of the MDOE models. Given the uniformly better quality that MDOE designs appear to deliver by every other metric considered in this study, this result may seem surprising. However, it is easily explained by consulting Eq. (21), which shows that the prediction uncertainty decreases as the square root of the number of data points used to fit the model. The OFAT design featured more than 10 times as much data as the MDOE design (with a parallel cost difference), and so would expect to deliver higher precision, especially in a random-error-only environment. While the MDOE design delivered less precision than the OFAT design, the precision was still adequate to satisfy quality standards, with both designs delivering uncertainty levels well below 0.25% of full scale output. This is an indication that the relatively large data volume of the OFAT design is not necessary to achieve quality objectives. MDOE designs achieve significant cost savings by "scaling" the experiment, to ensure that ample data are acquired to achieve precision goals, but no more.

J. Number of Type I Inference Errors

1. Description and Computation

The terms in the model describing the "true" balance responses are all known, since this is a simulation. We can therefore compare each math model to the true model to test for certain types of inference errors in constructing the regression models in the presence of a simulated noise environment. One of these is the Type I inference errors: We commit a Type I inference error when we erroneously reject a null hypothesis. In the regression process, there is an implicit null hypothesis for each term, stating that the term is not significant and does not belong in the model. We reject the null hypothesis when we include a given term in the model. If we erroneously reject the null hypothesis, it means that we have included a term in the model that does not belong there. Since the terms in the "true" model are known in this simulation study, Type I inference errors were easy to count by simply observing how many terms there were in each fitted response model that were not in the original, "true" model.

2. Observations

Observations for this model quality metric are drawn from Fig. 27.

J1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
J2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
J3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
J4. Adding the systematic component of unexplained variance increased the lack-of-fit F-statistic when the data were acquired with the OFAT experiment design, but not when the data were acquired with the MDOE experiment design.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the number of Type I coefficient inference errors for models obtained by DX7 and BALFIT. This conclusion is supported by observation J3.

- We reject H_{02}: B=0, concluding that MDOE experiment designs lead to a smaller number of Type I coefficient inference errors than OFAT experiment designs. This conclusion is supported by observations J1 and J2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to a greater number of Type I coefficient inference errors than when there is only random noise. This conclusion is supported by observations J1 and J2.
- We reject H_{14}: AB≠0, concluding that differences in the number of Type I coefficient inference errors resulting from MDOE and OFAT experiment designs is independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation J3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the number of Type I coefficient inference errors that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation J3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on the number of Type I coefficient inference errors, depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations J1 and J4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for number of Type I coefficient inference errors. This means that the significant interaction observed between DESIGN and NOISE for the number of Type I coefficient inference errors is the same whether the models are developed with BALFIT or DX7. This conclusion is supported by observation J3.

4. Discussion

Models developed by BALFIT and DX7 appear to be equally adept at excluding extraneous terms from their recommended models. As with other metrics examined in this study, for either software package the number of Type I coefficient inference errors is adversely impacted by the addition of systematic error for OFAT experiment designs. For random noise only, OFAT and MDOE models had a negligible number of Type I coefficient inference errors—an average too close to zero to distinguish it from zero. MDOE models also generated too few Type I inference errors in the presence of random-plus-systematic noise to distinguish the average from zero. However, the introduction of systematic error did increase the number of Type I inference errors for models developed with an OFAT design. OFAT models developed in the presence of systematic error produced an average of between one and two Type I inference errors. That is, these models typically had one or two terms in them that were not in the "true" model. For obvious reasons, this can introduce bias by adding fictitious components to the response prediction. Systematic variations in the data introduce another regressor that is not addressed in models based on OFAT experiment designs; namely, time. The least squares algorithm tries to compensate for this unknown factor by introducing loading cross-terms to account for it. The MDOE designs are randomized, however, which converts systematic variations to an additional component of random error that does not impact the shape of the response surface model. This is why the OFAT designs generate more Type I inference errors than the MDOE designs.

K. Number of Type II Inference Errors

1. Description and Computation

We commit a Type II inference error when we erroneously reject an alternative hypothesis. For each null hypothesis in the regression process, there is a corresponding alternative hypothesis for each term, stating that the term is significant and therefore belongs in the model. We reject the alternative hypothesis when we exclude a candidate term from the model. If we erroneously reject the alternative hypothesis, it means that we have left a term out of the model that really belongs there. As with the Type I inference errors (the error made by including terms that do not belong), it was easy to count Type II inference errors because this is a simulation for which the true model is known. Any term in the "true" model that was not included in the recommended model was counted as a Type II inference error.

2. Observations

Observations for this model quality metric are drawn from Fig. 28.

K1. The A (SOFTWARE), B (DESIGN), and C (NOISE) effects are all significant with respect to the t-Value Limit limit. The B and C effects are significant by the Bonferroni Limit.
K2. All three significant effects are positive.
K3. No interaction effect (AB, AC, BC, or ABC) is significant.

3. Conclusions

- We reject H_{11}: A≠0, concluding that there is no significant difference in the number of Type II coefficient inference errors for models obtained by DX7 and BALFIT. This conclusion is supported by observation K1. While the SOFTWARE effect is large enough to be regarded as significant at the t-Value limit (95% probability that this individual effect is significant), there is less than a 95% probability that this effect is also significant if the B and C terms are (the Bonferroni Limit). See Fig. 28b.
- **We reject H_{02}: B=0, concluding that MDOE experiment design lead to a larger number of Type II coefficient inference errors than the OFAT experiment design. This conclusion is supported by observations K1 and K2**.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to a greater number of Type II coefficient inference errors than when there is only random noise. This conclusion is supported by observations K1 and K2.
- We reject H_{14}: AB≠0, concluding that differences in the number of Type II coefficient inference errors resulting from MDOE and OFAT experiment designs is independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation K3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the number of Type II coefficient inference errors that are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation K3.
- We reject H_{16}: BC≠0, concluding that the addition of the systematic component of unexplained variance had the same effect on the number of Type II coefficient inference errors whether the data were acquired with the OFAT experiment design or the MDOE experiment design. This conclusion is supported by observations K3.
- We reject H_{17}: ABC≠0, concluding that there is no three-way interaction for number of Type II coefficient inference errors. This conclusion is supported by observation K3.

4. Discussion

For this specific study, BALFIT appeared to be very slightly more likely to produce models with Type II inference errors than DX7. That is, there is a slight tendency for BALFIT to overlook subtle effects in the true model. However, none of the terms from the true model that were excluded from the BALFIT recommended models was large enough to cause a practical difference in response predictions. Even absent these small terms, the accuracy of the BALFIT models was well within typical precision requirements for force balance calibration. The difference between the BALFIT and DX7 results was too small to be resolved at the Bonferroni Limit, and for this reason the null hypothesis of no significant software difference was not rejected.

Likewise, models produced from the MDOE design had more Type II inference errors than those generated from the OFAT design. This is attributed to the 10:1 data volume difference between the OFAT and MDOE designs that enabled the OFAT models to resolve terms in the true model that were real, but nonetheless too small to have any practical impact on the response predictions. That is, the OFAT models were largely fitting noise. The MDOE designs produced smaller models, which by Eq. (21) minimized the prediction uncertainty and produced a higher-quality result.

L. Number of Erroneously Estimated Coefficients

1. Description and Computation

There are three ways to make errors in constructing a regression model: We can include a term that does not belong in the model, we can exclude a term that does belong in the model, or we can improperly estimate the coefficient for a term that we have correctly included in the model. To estimate the latter error, we computed 95% confidence intervals for each individual regression coefficient and counted the number of times that the true model's corresponding coefficient was outside of this interval.

2. *Observations*

Observations for this model quality metric are drawn from Fig. 29:

L1. The B (DESIGN), C (NOISE), and BC effects are all significant with respect to the Bonferroni limit.
L2. The C (NOISE) effect is positive while the B (DESIGN) and BC effects are negative.
L3. Neither the A (SOFTWARE) effect nor its interaction with any other effects (AB, AC, or ABC) is significant.
L4. Adding the systematic component of unexplained variance increased the number of erroneously estimated coefficients significantly when the data were acquired with the OFAT experiment design, and much less so when the data were acquired using MDOE.

3. *Conclusions*

- We reject H_{11}: A≠0, concluding that there is no significant difference in the number of erroneously estimated coefficients for models obtained by DX7 and BALFIT. This conclusion is supported by observation L3.
- We reject H_{02}: B=0, concluding that MDOE experiment designs lead to a smaller number of erroneously estimated coefficients than OFAT experiment designs. This conclusion is supported by observations L1 and L2.
- We reject H_{03}: C=0, concluding that the addition of a systematic component to an existing random component of unexplained variance leads to a greater number of erroneously estimated coefficients than when there is only random noise. This conclusion is supported by observations L1 and L2.
- We reject H_{14}: AB≠0, concluding that differences in the number of erroneously estimated coefficients resulting from MDOE and OFAT experiment designs are independent of whether BALFIT or DX7 was used to generate the models. This conclusion is supported by observation L3.
- We reject H_{15}: AC≠0, concluding that the addition of a systematic component to an existing random component of unexplained variance causes differences in the number of erroneously estimated coefficients that are independent of whether BALFIT or DX7 was used to generate the models This conclusion is supported by observation L3.
- We reject H_{06}: BC=0, concluding that the addition of the systematic component of unexplained variance had a different effect on the number of erroneously estimated coefficients, depending on whether the data were acquired with the OFAT experiment design or the MDOE experiment design. By this metric, the quality of models developed with the MDOE design seems to be more robust with respect to the addition of systematic error than the quality of models developed with the OFAT design. This conclusion is supported by observations L1 and L4.
- We reject H_{17}: ABC≠0, concluding that the level of any two-way interaction is independent of the level of the third factor for number of erroneously estimated coefficients. This means that the significant interaction observed between DESIGN and NOISE for the number of erroneously estimated coefficients is the same whether the models are developed with BALFIT or DX7. This conclusion is supported by observation L3.

4. *Discussion*

Models developed by BALFIT and DX7 appear to perform equally well at estimating regression coefficient values. As with other metrics examined in this study, for either software package the number of erroneously estimated coefficients is adversely impacted by the addition of systematic error for OFAT experiment designs. Such errors increased from an average of one to three for random noise to an average of five to seven when both random and systematic errors were in play. The MDOE models, by comparison, did not generally produce coefficient errors under pure random error, and generated on average only one relatively small coefficient error in the presence of systematic error. This increase was too small to resolve with 95% confidence. Furthermore, the MDOE error tended to be in the intercept term and not in one of the terms regressors, meaning that the dependence of response on loading levels was preserved. This also meant that the response error was constant—independent of the levels of the loading variables and therefore easier to take into account.

M. R-Squared Statistics

1. Description and Computation

R-squared statistics are commonly used to assess the quality of regression models. Three variations were examined for each of the models in this study:

- **Ordinary R-Squared**: Ratio of the explained sum of squares to the corrected total sum of squares. The corrected total sum of squares is computed by adding all the squared differences between each simulated measured response and the average of all the simulated measurements. The *explained* sum of squares is computed by adding all the squared differences between each predicted response and the average of all the simulated measurements. The R-Squared statistic approaches 1 as the model explains more and more of the variability in the data.
- **Adjusted R-Squared:** Ratio of the total *explained* variance to the total variance. Computed the same way as the ordinary sum of squares, except that the numerator and denominator are each adjusted as follows: The explained sum of squares is first divided by the number of degrees of freedom associated with it, which is just the number of regressors in the model ($p - 1$). The result is the explained mean square, or explained variance. The corrected total sum of squares is likewise first divided by the corrected total degrees of freedom, which is just the number of points in the data set minus 1. The result is the total mean square, or variance. The adjusted R-Squared is often recommended as an improvement over the ordinary R-Squared statistic because the latter can be made arbitrarily close to one by simply adding regressors to the model. The Adjusted R-Squared tends to plateau as the number of regressors increases. As with the ordinary R-Squared statistic, the adjusted R-squared statistic approaches 1 as more of the total variance in the data is explained by the model.
- **Predicted R-Squared:** Computed by subtracting the ratio of the PRESS statistic to the total sum of squares from 1. The PRESS statistic (Predicted Residual Sum of Squares) is calculated by removing each data point from the regression in turn, computing the model coefficients based on the remaining n-1 data points, using this model to predict the response at that point, and forming a residual by subtracting that predicted response from the simulated measured response. The square of all such residuals is summed to form the PRESS statistic. The ordinary and adjusted R-Squared statistics purport to quantify the fraction of the total variability in the *existing* data set that can be explained by the model. The predicted R-squared statistic is a measure of how much variability in *new* data that the model is expected to explain.

2. Observations

Observations for this model quality metric are drawn from Fig. 30.

None of the three main factor effects—A (SOFTWARE), B (DESIGN), or C (NOISE)—was significant for any of the above variations of R-Squared, nor were any interactions among those factors significant.

3. Conclusions

We conclude the R-Squared statistics are poorly suited to provide insight into the effects of software, experiment design, and noise level on the quality of regression models.

4. Discussion

This result was somewhat unanticipated, given the ubiquitous use of R-squared statistics to validate model quality. R-Squared values were typically "1" to seven or more significant figures in this study, leaving very little to choose between the largest and smallest values. None of the conclusions in the next section rest on R-Squared measures.

N. Summary of Comparisons and Results of Analysis

The principal findings of this section are summarized in Table 13. Clearly the model-building process simulated in this study is dominated by experiment design and noise environment. No significant software effect is observed, nor does the choice of analysis software appear to impact the noise and design effects or their interaction.

Table 13. Impact of Experiment Design, Noise Environment, and Modeling Software on Selected Model Quality Metrics: Summary of Significant Effects. Filled circles mark effects that are Significant at the Bonferroni Limit. Open Circles are Significant at the t-Level only. Empty cells Identify Insignificant Effects.

Metric	Effects						
	A: SOFTWARE	B: DESIGN	C: NOISE	AB	AC	BC	ABC
Standard Deviation of Model Residuals		●	●			●	
Maximum Model Residual	○	●	●			●	
Standard Deviation of Confirmation Points		●	●			●	
Maximum Confirmation Point Residual		●	●			●	
Number of Successful Confirmations		●	●			●	
Lack of Fit F-Statistic		●	●			●	
Average Prediction Uncertainty		●	●			●	
Coefficient Type I Inference Errors		●	●			●	
Coefficient Type II Inference Errors	○	●	●				
Improperly Estimated Coefficients		●	●			●	
Ordinary R-Squared							
Adjusted R-Squared							
Predicted R-Squared							

VII. Summary and Conclusions

Simulated strain-gage balance calibration data have been used to assess differences in two balance calibration model building methods for different noise environments and experiment designs. One method uses a customized software system developed at Ames Research Center and the other employs standard response surface modeling methods implemented in numerous commercially available data analysis software products. For each of six simulated balance outputs, calibration models were developed for a total of eight combinations of experiment design, noise environment, and software system. A total of 48 math models were therefore developed and compared on the basis of a number of model quality assessment metrics. Overall, the performance of models developed with the two methods shows very good agreement. By a variety of model quality assessment metrics, no significant difference could be detected between models built with the two methods.

Experiment design and noise environment had a much greater impact on the quality of the final calibration models than the method used to develop the models. Calibration models developed using the Modern Design of Experiments were superior to models developed using conventional One Factor At a Time loading schedules, especially when there was a systematic component of the unexplained variance as happens when long-term persisting effects are in play during an experiment (thermal effects, instrument drift, etc). There was a pronounced interaction between experiment design and noise environment, with formally designed experiments having a greater impact on model quality under more imperfect noise conditions that unfortunately characterize many realistic measurement environments.

While the two software systems exhibited similar performance, the Ames software system is customized for balance calibration data analysis and therefore offers certain efficiencies with respect to commercial, off-the-shelf software. For example, certain calculations specific to balance calibration are automated in the Ames system. These include tare-load iterations, and also model inversions to express the load variables as a function of the response variables (as is ultimately necessary in order to use a balance to acquire force and moment data.) The Ames Software is also configured to account for absolute-value terms in the fitted model, which adds complexity to the analysis if performed by general-purpose software. Perhaps the greatest strength of the Ames software is its automated report-writing capability, which enables large amounts of useful information to be automatically cast in a

portable data format that enables widespread circulation. The commercial software systems are generally much more limited in this regard.

The general-purpose software features a larger number of model quality assessment metrics, including tests for multicollinearity. Multicollinearity is a state in which model regressors can become highly correlated due to a physical constraint involving the component loads that are applied in a balance calibration experiment using a Single Vector System. The general software also permits the automated imposition of hierarchy in the math models, as is necessary to render them invariant under certain translational transformations. Finally, the general software utilizes coded variables, which have certain computational and interpretive advantages. The authors have agreed to a collaboration intended to adopt the best features of both systems in an updated version of the Ames software, to be reported at the 2007 summer AIAA Joint Propulsion Conference.

Appendix A. Coding of Independent Variables and Hierarchy

A. Coding of Independent Variables

Coding is a procedure that invokes a linear transformation to map physical units into a dimensionless scale, typically from -1 to +1. Let ξ represent some independent variable such as the normal component of force in a balance load schedule, and let L and H represent the lowest and highest values of this variable in physical units, say, pounds. The following transformation converts a physical variable ranging from L = -6520 lbs to H = +6520 lbs, say (from Table 4), into a coded variable ranging from -1 to +1. This transformation both scales and centers the variable:

$$x = \frac{\xi - \tfrac{1}{2}(H + L)}{\tfrac{1}{2}(H - L)} \tag{A1}$$

An inverse transformation can be used to convert back to physical units, as follows:

$$\xi = \tfrac{1}{2}\left[H(x+1) - L(x-1)\right] \tag{A2}$$

The inverse transformation may be problematical unless hierarchy is maintained, as will be discussed presently.

There are many reasons for using coded variables in regression calculations:

Greater clarity: Coded variables can bring a certain level of clarity to the problem that is more difficult to achieve when physical units are used because it is easier to see the relative contributions of various terms in a model when the regression has been performed on coded variables rather than physical variables. For example, the principal load coefficient for the normal force coded model in Table 3 is 2096.7, which can be seem immediately to be three orders of magnitude larger than the next-most influential term in the model, the interaction between side force and rolling moment, which has a coded-variable coefficient of 7.08. Likewise, it is clear that most of the interaction terms in the normal force model have coefficients that are less than 1.0 in coded units, indicating clearly how small the interaction terms are. This, coupled with the fact that the quadratic principal load term is only 2.16, indicates at a glance the normal force output is a near linear function of its inputs. When the variables are in physical units, the numerical values of the coefficients depend on which units are selected, and the relative importance of some terms becomes less obvious.

Note also that the load range for the NF model of Table 3 is well approximated by the coefficient of the principal load term, as this coefficient is defined as the change in normal force response due to a change in coded variable level from 0 (no load) to 1 (full load). A more precise estimate is easy to obtain by also accounting for the coefficient of the quadratic term.

By centering the variable, coding ensures that zero is always within the range of every independent variable. This attaches a physical interpretation to the intercept; namely, that it is the average of the response measurements. The regression intercept does not always a physical meaning if the independent variable range does not include zero, which can be the case if the variables are not coded. Centering also decouples slope and intercept effects. If the independent variable range contains zero, changes in the slopes of a response function (its shape) are independent of changes in the mean level of responses (the intercept). This ensures that the functional form of the response model is decoupled from issues associated with precisely determining the intercept. Bias errors therefore affect only the intercept, and not the details of how responses depend on the independent variables.

Orthogonality: Consider a regression model with K regressors that is refitted to the data as a function of K-1 of the regressors. If the coefficients of all retained regressors remain unchanged, we say that the discarded regressor was *orthogonal* to the remaining regressors. Orthogonality is a desirable property because it ensures that the magnitude of a given coefficient is independent of additional terms that may or may not be in a model. This makes the interpretation of regressor effects independent of other terms in the model, which is helpful in understanding the underlying physics.

Consider the following simple illustration of a two-factor, two-level test matrix, where the variables are expressed in physical units (ξ_i) ranging from 1 to 10, and in coded units (x_i) ranging in the usual way from -1 to +1. Equations (A1) and (A2) can be used in this example to convert from physical to coded units and back by setting $L = 1$ and $H = 10$.

Table 14. Comparison of coded and physical units.

Point	Physical			Coded		
	ξ_1	ξ_2	$\boldsymbol{\xi_1\xi_2}$	x_1	x_2	$\mathbf{x_1x_2}$
1	1	1	**1**	-1	-1	**-1**
2	1	10	**10**	-1	+1	**+1**
3	10	1	**10**	+1	-1	**-1**
4	10	10	**100**	+1	+1	**+1**
		Sum:	**121 ≠ 0**		**Sum:**	**0**

Recall that two vectors are orthogonal if the sum of their term-by-term cross-products (proportional to the cosine of the angle between the two vectors) is zero. Since a test matrix is comprised of a set of vectors representing independent variable levels, the same test of orthogonality can be applied to show that x_1 and x_2, are orthogonal but ξ_1 and ξ_2 are not. This means that if a model is fitted in terms of physical units, the value of the ξ_1 coefficient will depend upon whether ξ_2 is in the model or not (and conversely), but if the model is fitted in terms of coded units, the value of the x_1 term will be the same whether x_2 is retained or not, and conversely. This is especially relevant in the current context of model reduction by the elimination of terms. When the regressors are not orthogonal, each estimated regression coefficient is in fact a function of the true coefficients of more than one regressor. The least-squares algorithm commonly used in regression computations is designed to determine the set of coefficients that minimizes the residual sum of squares, and will often produce relatively small test matrix residuals even when the regressors are not entirely orthogonal. However, if the regressors would have had different values with a different set of terms in the model, the coefficients that minimize residual error for the design matrix points may not predict responses at other points with the least possible error. A class of functions known as orthogonal polynomials has been used in aerospace response surface modeling applications to predict responses using math models for which every term is orthogonal to all other terms.[24] But absent the use of such a specialized class of functions, orthogonality is difficult to eliminate entirely in general regression applications even when the variables are coded. Coding can provides some incremental quality improvement, however.

Computing Resolution: Real numbers of the kind used in regression computations are a mathematical abstraction involving infinite resolution and infinite range. Computers must approximate real number calculations by using floating point numbers that involve a finite set of values with finite precision. The ANSI/IEEE Standard 754-1985 for Binary Floating-Point Arithmetic[25] establishes conventions for double-precision floating point numbers that specify they be stored in 64-bit words, with 52 bits reserved for a mantissa, 11 bits reserved for an exponent, and one bit to represent the sign. The finite bit allocation for the exponent limits the range of floating point numbers to something on the order of $\pm 10^{308}$, and the finite number of mantissa bits limits the resolution to $2^{-52} = 2.22\text{E-}16$ at best. (Actually, the mantissa and exponent interact in such a way that the spacing between realizable floating point numbers gets larger for larger numbers). The range limitation does not generally represent a practical constraint in regression analysis but the resolution limit can affect calculations. Consider this example:

It can be shown that the vector of regression coefficients from Eq. (1b), β, can be computed as follows:

$$\beta = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{A3}$$

where **X** is the design matrix and **y** is the vector of observed responses.

To illustrate the point about floating point resolution, consider a simple example in which the design matrix is as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & \delta \\ 1 & \delta & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{A4}$$

This design matrix corresponds to a first-order math model in two variables, to be fitted with three data points. We then have

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & \delta & 0 \\ \delta & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & \delta \\ 1 & \delta & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & \delta & \delta \\ \delta & \delta^2 & 0 \\ \delta & 0 & \delta^2 \end{bmatrix} \tag{A5}$$

If $|\delta| < 10^{-8}$, a floating point representation of this matrix is singular even with double-precision, because the resolution limit of 10^{-16} dictated by the IEEE standard renders the 2nd and 3rd columns identical. The inverse would then be nonexistent and the regression coefficients could not be computed by Eq. (A3). Depending on the details of the test, it is not inconceivable that this limit could be encountered when the variables are expressed in physical units. For example, simply measuring force balance outputs in volts instead of microvolts could produce very nearly singular results when the calibration equations are inverted to express loads as a function of electrical output levels. A high correlation between regressors can result in a model that does not predict responses for arbitrary independent variable combinations as well as it does for the test matrix points for which the model coefficients were optimized. Coding the independent variables circumvents these floating point resolution issues and minimizes the potential for this source of multicollinearity.

B. Hierarchy

Peixoto[26] describes a general coding transformation consisting of a combination of scaling and translation, and demonstrates that while *scaling* transformations do not change the estimation space of a polynomial function of two or more variables, the estimation space of such a polynomial is invariant under *translation* if and only if the polynomial is well-formulated. Following Kempthorpe,[27] he uses the term "well formulated" to refer to a polynomial in which no hierarchically inferior terms have been eliminated.

Such a model contains all the components of terms that are second-order and higher. For example, if an AB interaction term is included in the model, both the A and the B first-order terms must also be included to maintain hierarchy. Likewise, if terms of the form A^2B are in the model, then so must A, B, and AB for the model to be well-formulated.

Note that Eqs. (A1) and (A2) represent coding transformations that include both scaling and translation, so it is important to maintain hierarchy when the variables are coded in this way. This means that hierarchically inferior terms must be retained in order to maintain hierarchy and preserve the invariance of the coding transformation, even if they are statistically insignificant. This is especially relevant when some multicollinearity is present, so that the coefficients are not all independent. In such a case, the coefficient of a term of second order or higher could be a function of the coefficients of its hierarchically inferior components, so that removing them from the model skews the coefficient of some of the retained terms.

Also, the residual sum of squares of a nonhierarchical model includes components due to the missing hierarchically inferior terms. This makes the residual variance less representative of the true experimental error. For this reason, some commercial packages (e.g., Minitab®[18]) will not perform an analysis of variance on nonhierarchical models. Design Expert[17] will permit such an analysis, but only under duress. Whenever a model is generated with missing hierarchically inferior terms, Design Expert generates a prompt to permit hierarchy to be automatically reinstated at the user's option. If the user declines, Design Expert provides a warning and a second

prompt. If the user still declines the option to make the model hierarchical, Design Expert generates the model, but with this disclaimer:

> "Using this non-hierarchical polynomial regression model (it excludes hierarchically inferior terms) is not recommended. Measures of goodness of fit and the predicted response values may be not be the same as those from the coded equation. All analysis within Design-Expert software is based on the coded equation."

Draper and Smith[28] also argue against dropping hierarchically inferior terms under a translation of origin. They propose the following rule:

> "If a model is to be consistent under a shift in origin, only the highest-order terms can be deleted at first and any chosen deletions must keep the model well-formulated. Moreover, if any of the highest-order terms are retained, all terms of lower order affected by them in a shift of origin must also be retained, whether or not their estimates are significant in the regression fit."

Draper and Smith also provides guidelines for removing terms when a rotational transformation is applied, as is commonly the case when certain canonical forms are invoked in the analysis of response surface models in order to remove cross-terms from the model. This is useful when it is of interest to determine variable combinations that correspond to a response maxima or minima, for example.

While canonical analysis is not commonly invoked in the analysis of calibration data, it is important to note that even seemingly innocuous variable transformations can have unintended consequences that must be thoroughly understood. For example, Peixoto's original work on this subject was motivated by a study of average daily temperatures in 56 U.S. cities as a function of longitude and latitude in which he simply translated the origin of the longitudinal variables to center them in the United States. This ostensibly benign change was the equivalent of redefining the origin of longitude measurements to pass somewhere near St. Louis rather than through Greenwich, a change that would hardly be expected to influence temperature predictions. Nonetheless, the functional form of non-hierarchical third-order polynomial functions of longitude did change under this origin translation, while the functional form of hierarchical models remained the same.

Peixoto's temperature modeling example illustrates that nonhierarchical models are at a disadvantage under any translation of the origin, not just transformations used to code the variables. For example, it is common in balance calibration data to adjust the origin in various ways to account for tare loads. Models based on such data may be suspect if hierarchy is not maintained, even if the calculations are carried out in original physical units without coding the variables. As parting advice on this topic, Draper and Smith suggest that models generated by automated selection procedures should be reviewed and refined to ensure that they are well-formulated by the above criteria.

Appendix B: Multicollinearity

Regression models developed during the analysis of calibration data are intended to serve as general prediction tools, providing accurate estimates of balance outputs for any combination of loads within the range of the calibration. The individual regression coefficients also provide useful insights into the linearity of the balance, indicating which combinations of applied loads have the greatest effect on specific component responses. As has been described earlier in this paper, the regression coefficients also dictate which terms are retained in the final recommended model, with those having larger coefficients more likely to be retained than those with smaller coefficients. All of these purposes can be well served if there are no linear relationships among the regressors, in which case they are said to be orthogonal. Unfortunately, true orthogonality is in most practical applications of regression a mathematical abstraction that is only approximated at best by the regressors. Even two regressors that are totally unrelated tend not to be perfectly orthogonal, simply because of the presence of experimental error in the data used to estimate them.

Notwithstanding the infrequency of perfect orthogonality in general multivariate regression analysis, the essential goals of the regression can usually be adequately met unless there are near-linear dependencies among the regressors, in which case the model is said to possess an undesirable property known as *multicollinearity*. Unfortunately, this property is not as rare as one might hope in general aerospace research applications of regression analysis. One example where it can come into play is in the analysis of balance calibration data acquired with a single vector system (SVS) of loading.

SVS loading applies a single force vector to an offset from the balance moment center. The force is applied by hanging calibrated weights as in a conventional calibration loading system, with the orientation of the balance chosen to achieve the desired combination of component loads. This technique was introduced at Langley Research

Center to facilitate the implementation of MDOE experiment designs requiring multi-component load combinations that are impractical to set with conventional dead-weight systems.[6,7]

There is an inherent constraint in single-vector loading that imposed by the fact that the total moment vector so generated is always at right angles to the total force vector.[35] This results in a physical constraint involving three specific two-way interaction terms in a balance calibration model that renders them linearly dependent, as follows.

Consider a balance loaded with three components of force and three moment components. Let **F** be the total force vector acting on the balance and let **M** be the total moment vector. Using *x*, *y*, and *z* subscripts to denote the vector components, we can express the dot product of these two vectors as follows:

$$\mathbf{F} \bullet \mathbf{M} = \mathbf{F}_x\mathbf{M}_x + \mathbf{F}_y\mathbf{M}_y + \mathbf{F}_z\mathbf{M}_z = FM\cos(\theta) \tag{B1}$$

where F and M are magnitudes of the **F** and **M** vectors, respectively, and θ is the angle between them. For a SVS loading system, $\theta = \pi/2$ and the right side of Eq. (B1) is therefore zero.

Figure 31 displays the coordinate system for balance forces and moments recommended by the AIAA.[1] The arrows point in the direction of positive forces and moments. By this convention we can rewrite Eq. (B1) as follows:

$$-(\mathrm{YM})(\mathrm{NF}) + (\mathrm{PM})(\mathrm{SF}) - (\mathrm{RM})(\mathrm{AF}) = 0 \tag{B2}$$

The negative signs for the first and third terms result from a convention in North American wind tunnel testing in which the normal force is positive up and the axial force is positive downstream, which is opposite of the positive Z and X directions for the balance axis system. In any case, the three two-way interactions of Eq. (B2) are constrained, and this constraint results in near perfect correlation among the three terms that generates substantial multicollinearity when all three are retained in a model. This is because the constraint consumes one degree of freedom and since there are only two degrees of freedom remaining for these three terms, unique regression coefficients can only be estimated for at most two of them.

All MDOE calibration load schedules are designed at Langley Research Center to be implemented using SVS hardware. In practice this means that the optimum MDOE design, which distributes points in the six-dimensional loading space of the balance in order to achieve certain quality objectives (low uncertainty in estimates of the coefficients, high orthogonality, high prediction accuracy, etc.) must be slightly detuned to accommodate the constraint in Eq. (B2). (MDOE designs implemented by other than SVS dead-weight loading, such as by automated balance calibration machines, do not have to incorporate the **F•M** constraint.) The MDOE design simulated in this study generated loading combinations that were each subject to the constraint in Eq. (B2). Any one or two of these two-way interactions could be accommodated in a response model without injecting multicollinearity problems, but not all three.

If proactive steps are not taken to eliminate multicollinearity, it can have an adverse impact on the estimation of regression coefficients, as can be demonstrated by considering a simple example in which there are only two regressors, x_1 and x_2, scaled to unit length. The least-squares normal equations follow directly from Eq. (1b):

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y} \tag{B3}$$

or

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix} \tag{B4}$$

where r_{12} is the coefficient of correlation between x_1 and x_2, r_{jy} is the coefficient of correlation between the y and x_j, and the b_i are estimated regression coefficients.

The covariance matrix, Eq. (2), can be written for this example as follows:

$$\mathbf{C} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2 = \begin{bmatrix} \dfrac{1}{\left(1-r_{12}^2\right)} & \dfrac{-r_{12}}{\left(1-r_{12}^2\right)} \\ \dfrac{-r_{12}}{\left(1-r_{12}^2\right)} & \dfrac{1}{\left(1-r_{12}^2\right)} \end{bmatrix}\sigma^2 \tag{B5}$$

Recall that the diagonal elements of the covariance matrix represent the variance in least-squares estimates of the regression coefficients. If we extend this example to multiple regressors instead of two, the diagonal elements can be generalized as follows:

$$C_{jj} = \frac{\sigma^2}{1-R_j^2}, \quad \mathrm{j} = 1, 2, \ldots, p \tag{B6}$$

where p is the number of regressors in the model and the numerator contains the coefficient of multiple determination from regressing x_j on the remaining p-1 regressors.

The term Variance Inflation Factor,[36] or VIF, is used to describe the multiplier of σ^2 in Eq. (B6):

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2}, \quad \mathrm{j} = 1, 2, \ldots, p \tag{B7}$$

If a given regressor is highly correlated with any subset of other regressors in the model, the numerator of Eq. (B7) will become very small and the VIF will be very large. Table 15a illustrates how large the VIF values can become when multicollinearity is present.

All three terms of the dot product between the total force load vector and the total moment load vector were statistically significant for the models in Table 15a. Because of constraint Eq. (B2), these three regressors would be perfectly collinear except for the presence of some small experimental error, with theoretically infinite VIF values. Actual estimated VIF values, while not infinite, are sufficiently large to unambiguously indicate the presence of massive multicollinearity in these models. While perfect orthogonality implies a VIF of 1, it is a common convention to regard only VIF values greater than 10 as indicative of serious multicollinearity, although a more conservative criterion suggests that VIF values in excess of 5 are troublesome. Clearly all the terms in the four models presented in Table 15 are very nearly orthogonal except the three components of the **F•M** dot product.

Because high VIF values are associated with large variance in the estimates of the regression coefficients, different data samples featuring slightly different levels of the independent variables can produce widely varying estimates of the regression coefficients when multicollinearity is present, clearly an undesirable result. Also, since precision intervals for estimates of the regression coefficients are directly proportional to the square root of the corresponding diagonal element of the covariance matrix, the square root of VIF represents the factor by which uncertainty estimates for the regression coefficients are "inflated" due to multicollinearity (hence the name).

When there is near perfect colinearity as when a physical constraint is in play, a literal interpretation of the VIF value such as those highlighted in Table 14a is meaningless. In such cases the correct interpretation is to regard the VIF values as "infinite," meaning that the variance in estimates of the associated coefficients is so inflated that the coefficients can be virtually anything. That is, with perfectly collinear regressors there are simply too few degrees of freedom available to meaningfully quantify the uncertainty in estimates of the regression coefficients.

Table 15a. Variance Inflation Factors (VIFs) for Four Models that Retain All Three Components of F•M.

Regressor	Pitching Moment		Yawing Moment	
	Data Set 3	Data Set 4	Data Set 3	Data Set 4
NF	1.00	1.00	1.02	1.01
PM	1.01	1.01	1.02	1.02
YM	1.13	1.13	1.31	1.30
SF	1.01	1.01	1.01	1.01
(NF)(PM)	1.02	1.02		
(NF)(RM)			2.67	2.65
(NF)(YM)	**13,941,260**	**13,874,586**	**35,599,843**	**35,049,871**
(AF)(RM)	**4,340,756**	**4,319,998**	**11,085,246**	**10,913,998**
(SF)(PM)	**7,978,338**	**7,940,169**	**20,372,703**	**20,057,928**
(RM)(YM)	1.02	1.02		
(RM)(SF)	1.04	1.04		
(PM)(RM)			1.01	1.01
(NF)(NF)	1.01		1.02	

When one or more of the three terms in the **F•M** dot product are rejected during the regular process for determining the recommended model, the dot product constraint does not come into play and there are sufficient degrees of freedom available to estimate the significant terms in the model. However, if all three terms are initially retained in the model, it is possible to correct the multicollinearity problem by eliminating one of them. Subject matter expertise concerning the balance construction and other details may suggest one of the interaction terms to reject. If no such expertise is available, all combinations of one and two terms can be eliminated from the model in succession, with one or more model quality assessment metrics computed for each iteration. The configuration that generates the optimum set of quality metrics might then retained in the recommended model.

Table 15b presents Variance Inflation Factors for the same models as in Table 15a, except that the interaction between normal force and yawing moment is excluded. It illustrates the dramatic reduction in VIF levels that can be archived by simply dropping one of the collinear terms. Note that there is always a risk that the dropped term might have offered considerable explanatory potential, so that the model may now feature greater lack of fit, a circumstance that the model-builder should examine closely. Whatever other imperfections the Table 15b models may or may not have, they all are now highly orthogonal.

Table 15b. Variance Inflation Factors (VIFs) After Dropping the (NF)(YM) Component of F•M.

Regressor	**Pitching Moment**		**Yawing Moment**	
	Data Set 3	**Data Set 4**	**Data Set 3**	**Data Set 4**
NF	1.00	1.00	1.00	1.00
PM	1.01	1.01	1.01	1.01
YM	1.01	1.01	1.01	1.01
SF	1.00	1.00	1.00	1.00
(NF)(PM)	1.01	1.01		
(NF)(RM)			1.01	1.01
(NF)(YM)	**Deleted**	**Deleted**	**Deleted**	**Deleted**
(AF)(RM)	**1.04**	**1.04**	**1.04**	**1.04**
(SF)(PM)	**1.04**	**1.04**	**1.03**	**1.03**
(RM)(YM)	1.01	1.01		
(RM)(SF)	1.02	1.02		
(PM)(RM)			1.01	1.01
(NF)(NF)	1.01		1.00	

Multicollinearity limits the "transferability" of balance calibration regression results. The general problem of transferability is described succinctly in this quote from the AIAA Recommended Practice on Calibration and Use of Internal Strain-Gage Balances with Application to Wind Tunnel Testing[1]:

> "Typically, when a calibration matrix is applied to the same data in which it was derived, it produces a significantly lower standard deviation than its application to a set of data that contains different combinations of the independent variables. This lack of transferability indicates that the coefficients are biased toward the specific combinations of independent variables contained in the design that was used to generate the coefficients. Therefore, the estimates of standard deviation are not representative of the ability of the mathematical model to predict unknown loads throughout the entire six dimensional inference space."

The reason for the transferability problem is not hard to understand. The least-squares algorithm used to fit regression models distributes the coefficients in such a way as to optimize the fit for the specific points used to generate the model. In that sense the regression coefficients are "tuned" for a specific data set. In most practical circumstances, the fit will be at least as good for those points as for any other set, notwithstanding the fact that perfectly adequate response predictions can be made for other independent variable combinations with a well-formulated model.

Models suffer from the transferability problem more when their regression coefficients are estimated with a variance that is inflated due to multicollinearity, as Fig. 32 illustrates. This figure compares near-orthogonal pitching moment and yawing moment models from Data Set 3 of this study to models generated from the same data in which all three **F•M** components were significant. This induced a level of multicollinearity that is reflected in the VIF values of Table 15a.

The standard deviation of residuals and the largest residual are compared for two sets of data: the calibration data points used to fit the model, and an independent set of 25 confirmation points. The confirmation points were not used in the regression analysis that generated the model, but were held in reserve simply to test the model.

The transferability problem is revealed by the fact that for each pair of bars in Fig. 32, the right bar representing the confirmation point data is consistently higher than the left bar representing the model data. This difference is much greater when multicollinearity is present than when it is not. For the models displaying multicollinearity, the residual standard deviations and maximum residuals are almost an order of magnitude greater for confirmation-point data than the calibration data used to fit the models.

Note that the fitted model residuals are virtually the same whether multicollinearity is present or not; it is the confirmation-point residuals that show the effects of correlated regressors. This again reflects the fact that the least-squares algorithm minimizes residuals, whether regressors are highly correlated or not. This suggests that model residuals may provide an overly optimistic indication of the health of a response model. Whenever possible (and often it is not possible when data are provided by third parties for analysis), the authors recommend the acquisition of confirmation points in a calibration data set.

Acknowledgements

The authors wish to acknowledge the cooperation of Dr. Peter A. Parker of Langley Research Center for his assistance in providing information upon which the balance simulation was based. In addition, the first author acknowledges the support of the Langley Wind Tunnel Enterprise. The second author was supported by NASA Ames Research Center under contract NNA04BA85C.

References

[1]Recommended Practice: Calibration and Use of Internal Strain-Gage Balances with Application to Wind Tunnel Testing. AIAA R-091-2003.

[2]Cook, T.A., "A Note on the Calibration of Strain Gauge Balances for Wind Tunnel Models, Royal Aircraft Establishment (Bedford)," Technical Note No. AERO.2631, December 1959.

[3]Hansen, R.M., "Evaluation and Calibration of Wire-Strain-Gage Wind-Tunnel Balances Under Load," NACA Langley Aeronautical Laboratory, 1956.

[4]DeLoach, R., "Impact of Loading Selection and Sequencing on a Force Balance Calibration (Invited)," *AIAA 2001-0168*, 25th Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[5]DeLoach, R., "Tailoring Wind Tunnel Data Volume Requirements through the Formal Design of Experiments," *AIAA 98-2884*, 20th AIAA Advanced Measurement and Ground Testing Technology Conference, Albuquerque, New Mexico, June 1998.

[6]Parker, P., and DeLoach, R., "Response Surface Methods for Force Balance Calibration Modeling," 19th International Congress on Instrumentation in Aerospace Simulation Facilities, Cleveland, Ohio, August 2001.

[7]Parker, P.A., Morton, M., Draper, N., Line, W., "A Single-Vector Force Calibration Method Featuring the Modern Design of Experiments," *AIAA 2001-0170*, 39th Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[8]DeLoach, R., and Philipsen, I., "Stepwise Regression Analysis of MDOE Balance Calibration Data Acquired at DNW," *AIAA 2007-0144*, 45th Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2007.

[9]Philipsen, I. and Zhai, J., "Comparative Study of Strain-Gauge Balance Calibration Procedures Using the Balance Calibration Machine," *AIAA-2007-0143*, 45th Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2007.

[10]Parker, P.A.; and Rhew R.D., "A Study of Automatic Balance Calibration System Capabilities," Second International Symposium on Strain-Gauge Balances, Bedford, England, UK, May 1999.

[11]Montgomery, D.C., and Peck, E.A., *Introduction to Linear Regression Analysis,* 2nd ed., John Wiley and Sons, New York, 1992.

[12]Box, G.E.P., and Draper, N., *Empirical Model-Building and Response Surface,* John Wiley and Sons, New York, 1987.

[13]Ulbrich, N., and Volden, T., "Strain-Gage Balance Calibration Analysis Using Automatically Selected Math Models," *AIAA 2005-4084*, 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, Arizona, July 2005.

[14]Ulbrich, N., and Volden, T., "A New Approach to Strain-Gage Balance Calibration Analysis," 5th International Symposium on Strain-Gauge Balances, Aussois, France, May 2006.

[15]Ulbrich, N., and Volden, T., "Development of a New Software Tool for Balance Calibration Analysis," *AIAA 2006-3434*, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[16]Myers, R. H., and Montgomery, D. C., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, 2nd ed., John Wiley and Sons, New York, 2002.

[17]Design Expert, Software Package, Ver 7.03, StatEase, Inc., Minneapolis, Minnesota, 2006.

[18]Minitab, Software Package, Ver 14.2, Minitab Inc., State College, Pennsylvania, 2003.

[19]JMP, Software Package, Ver 6.0.3, SAS Institute, Cary, North Carolina, 2006.

[20]Statistica, Software Package, Ver 7.1, StatSoft, Inc., Tulsa, Oklahoma, 2006.

[21]MATLAB, Software Package, Ver 7.0 (R14), The Mathworks, Inc., Natick, Massachusetts, 2004.

[22]Ulbrich, N., and Volden, T., "Application of a New Calibration Analysis Process to the MK–III–C Balance," *AIAA 2006-0517*, 44th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2006.

[23]Ulbrich, N., and Volden, T., "Analysis of Floor Balance Calibration Data using Automatically Generated Math Models," *AIAA 2006-3437*, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[24]Morelli, E.A., and DeLoach, R., "Response Surface Modeling Using Multivariate Orthogonal Functions (Invited)", *AIAA 2001-0168*, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.

[25]IEEE Standard for Binary Floating-Point Arithmetic, ANSI/IEEE Std 754-1985, August 12, 1985.

[26]Peixoto, J.L., "A property of Well-Formed Polynomial Regression Models," *The American Statistician*, Vol. 44, No. 1, February 1990.

[27]Kempthorne, O., "Classificatory Data Structures and Associated Linear Models," *Statistics and Probability: Essays in Honor of C. R. Rao*, G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, Eds., Amsterdam, North Holland, pp. 397–410.

[28]Draper, N. R., and Smith, H., *Applied Regression Analysis*, 3rd ed., John Wiley and Sons, New York, 1998.

[29]Scheffe, H., *The Analysis of Variance*, John Wiley and Sons, New York, 1959.

[30]DeLoach, R., "Improved Quality in Aerospace Testing Through the Modern Design of Experiments (Invited)," *AIAA 2000-082,* 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2000.

[31]DeLoach, R., "Tactical Defenses Against Systematic Variation in Wind Tunnel Testing," *AIAA 2002-0885*, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2002.

[32]DeLoach, R., "Impact of Systematic Unexplained Variance on a Balance Calibration," 25th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.

[33]Box, G. E. P., and Cox, D. R., "An Analysis of Transformations," *J. R. Statist. Soc. Ser. B*, **26**, pp. 211–243, 1964.

[34]Coleman, H. W., and Steele, W. G., *Experimentation and Uncertainty Analysis for Engineers*, John Wiley and Sons, New York, 1989.

[35]Beer, P. P., and Johnston, E. R. Jr., *Vector Mechanics for Engineers,* McGraw-Hill, 1962.

[36]Marquardt, D. W., "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation," *Technometrics*, **12**, 591–612.

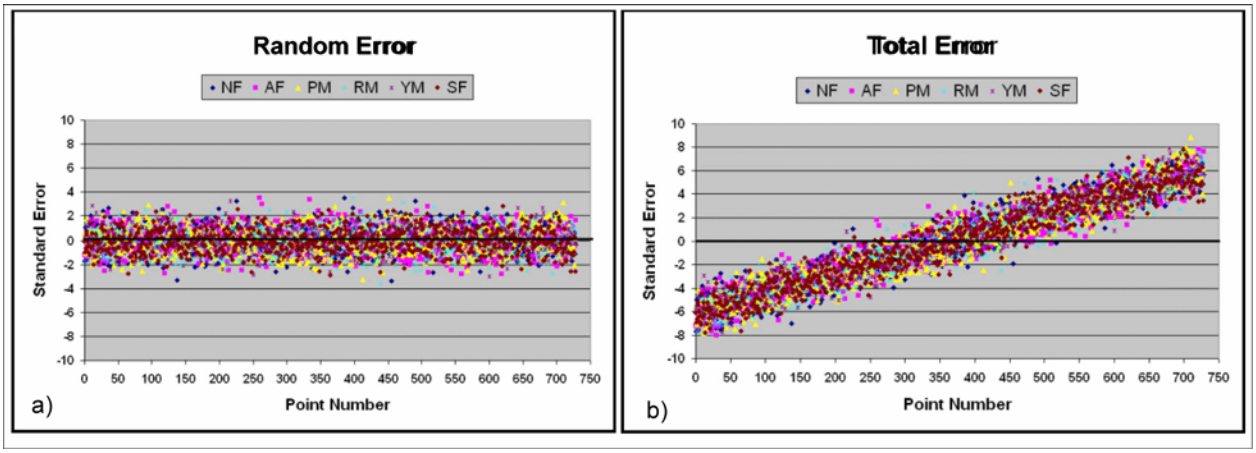

Figure 1. a) Random error, normalized by standard deviation; b) Random plus systematic error, normalized by standard deviation.

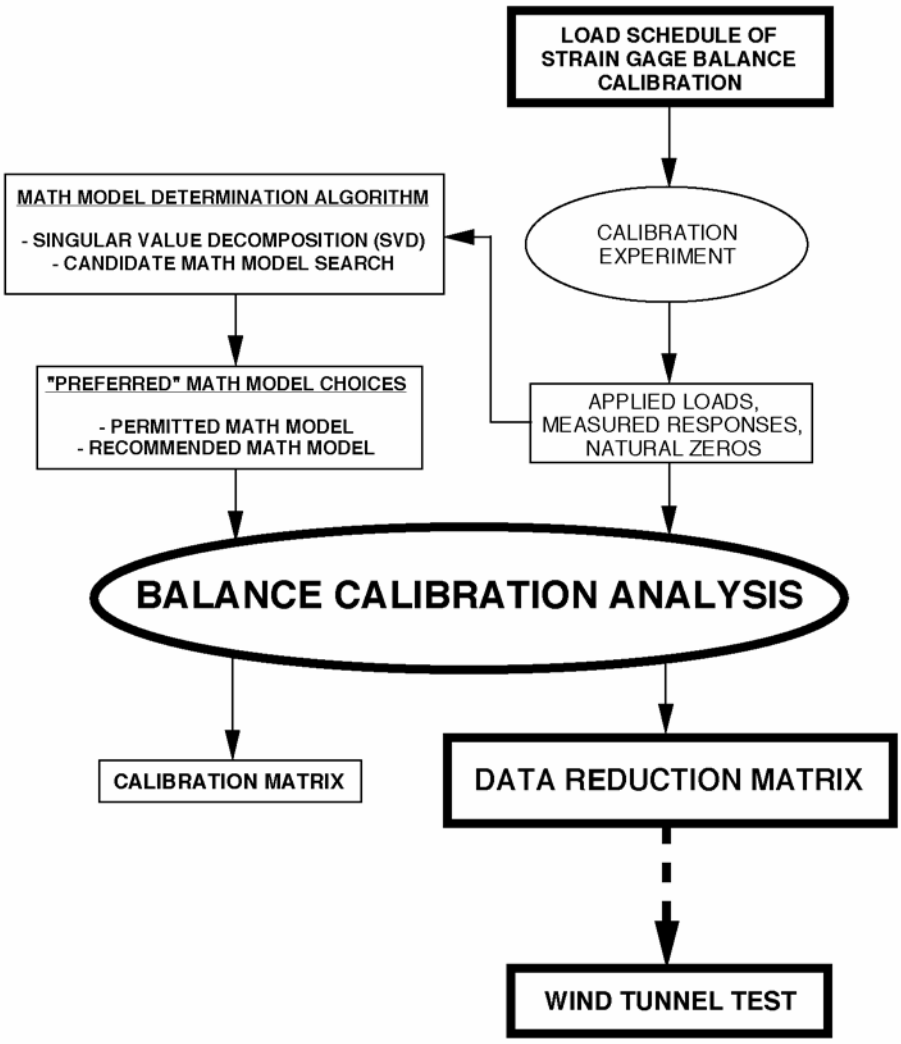

Figure 2. Key elements of the new Ames approach to strain-gage balance calibration analysis (BALFIT result).

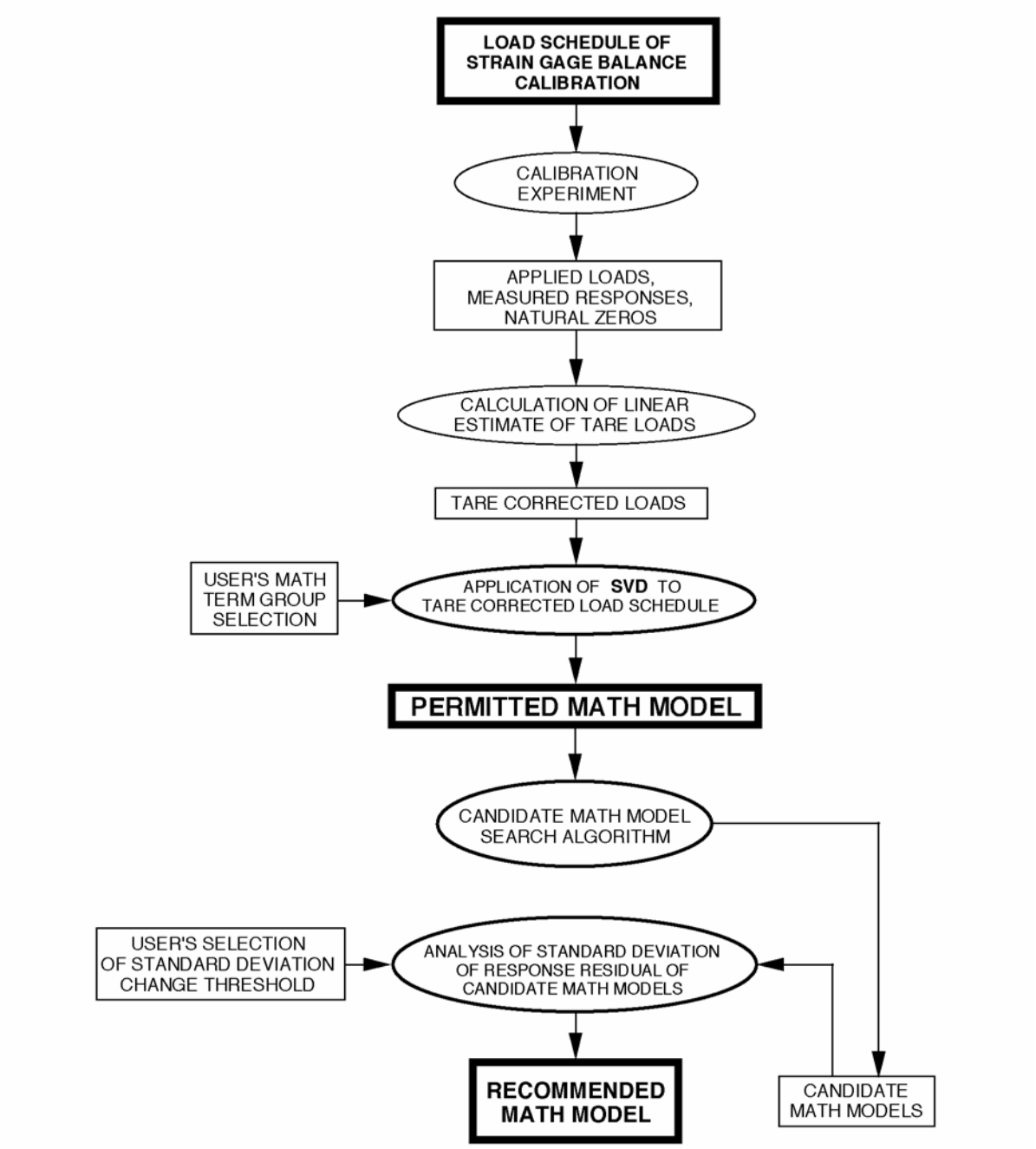

Figure 3. Determination of the permitted and recommended math model for balance calibration analysis (BALFIT result).

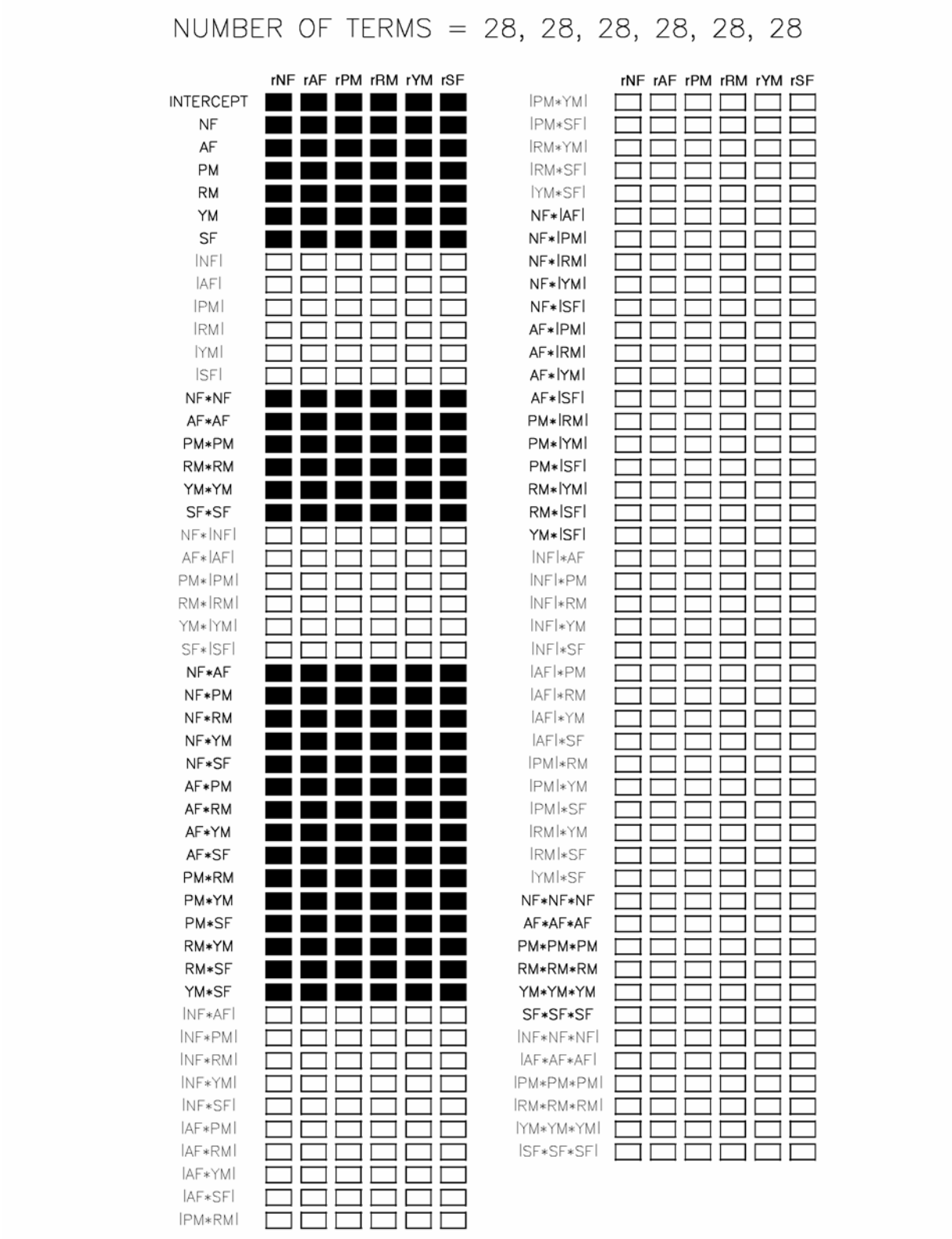

Figure 4a. Permitted math model used for analysis of Data Set 1 (BALFIT result).

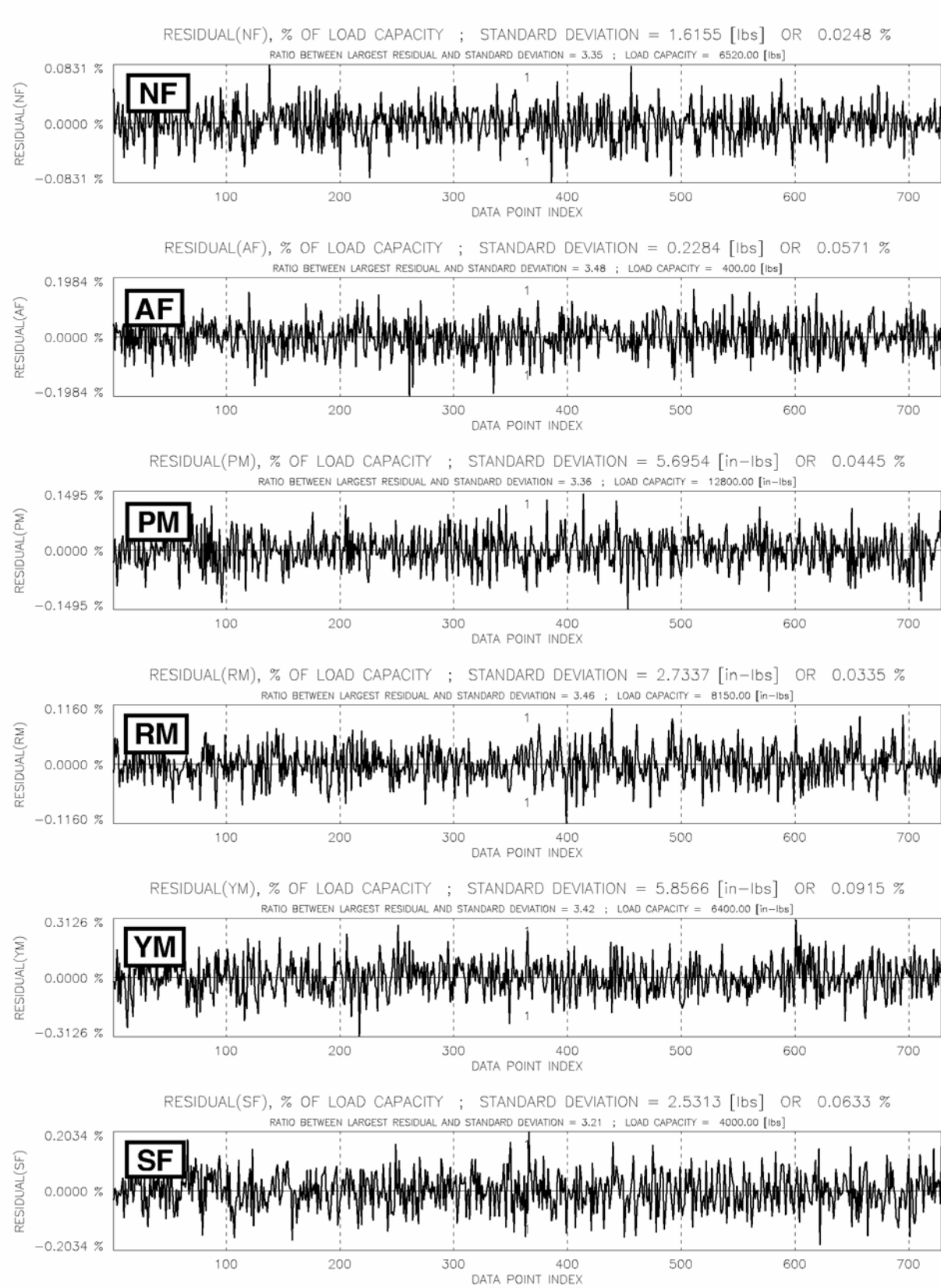

Figure 4b. Load residuals for permitted math model after analysis of Data Set 1 (BALFIT result).

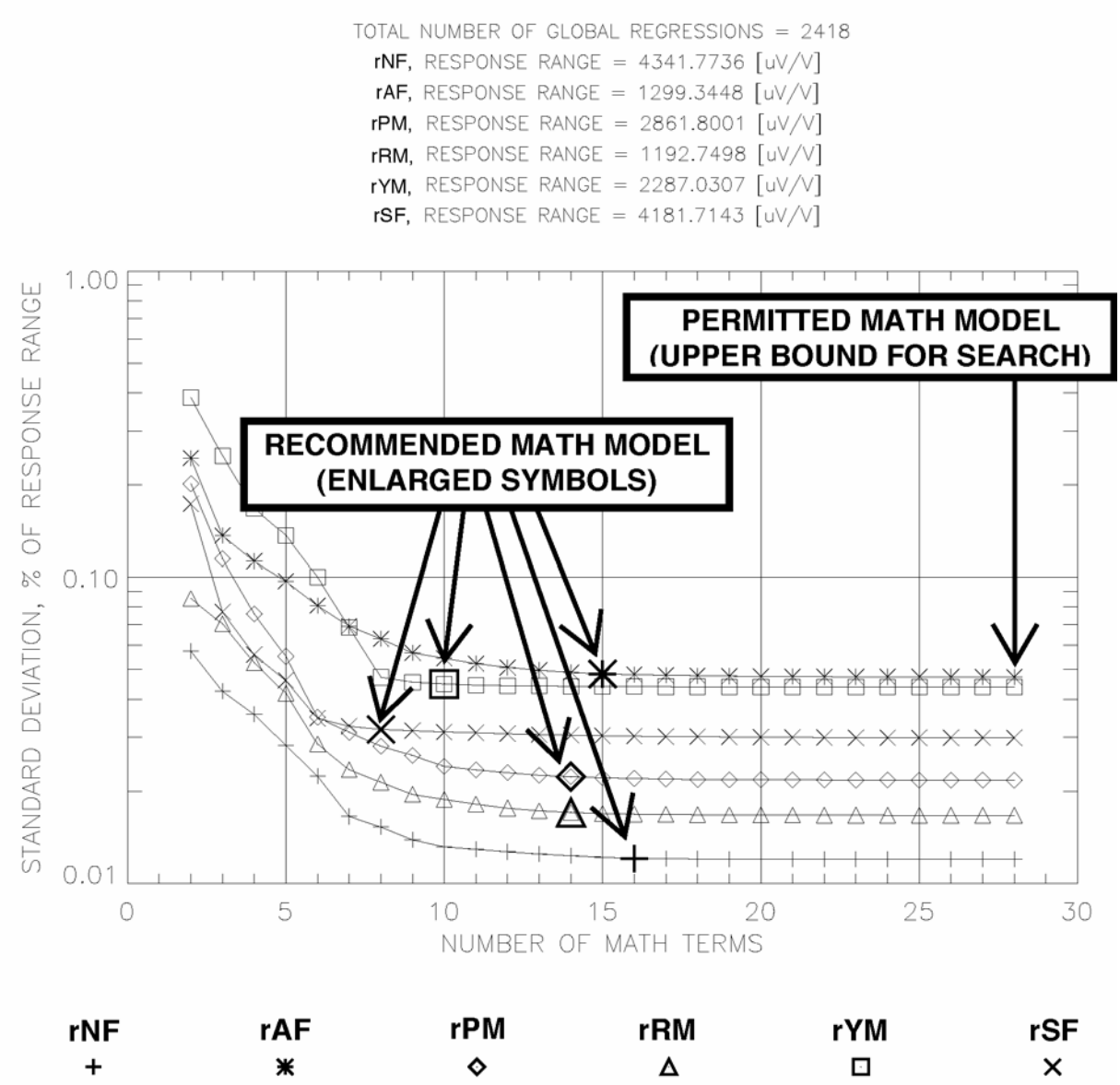

Figure 4c. Results of application of candidate math model search to Data Set 1 (BALFIT result).

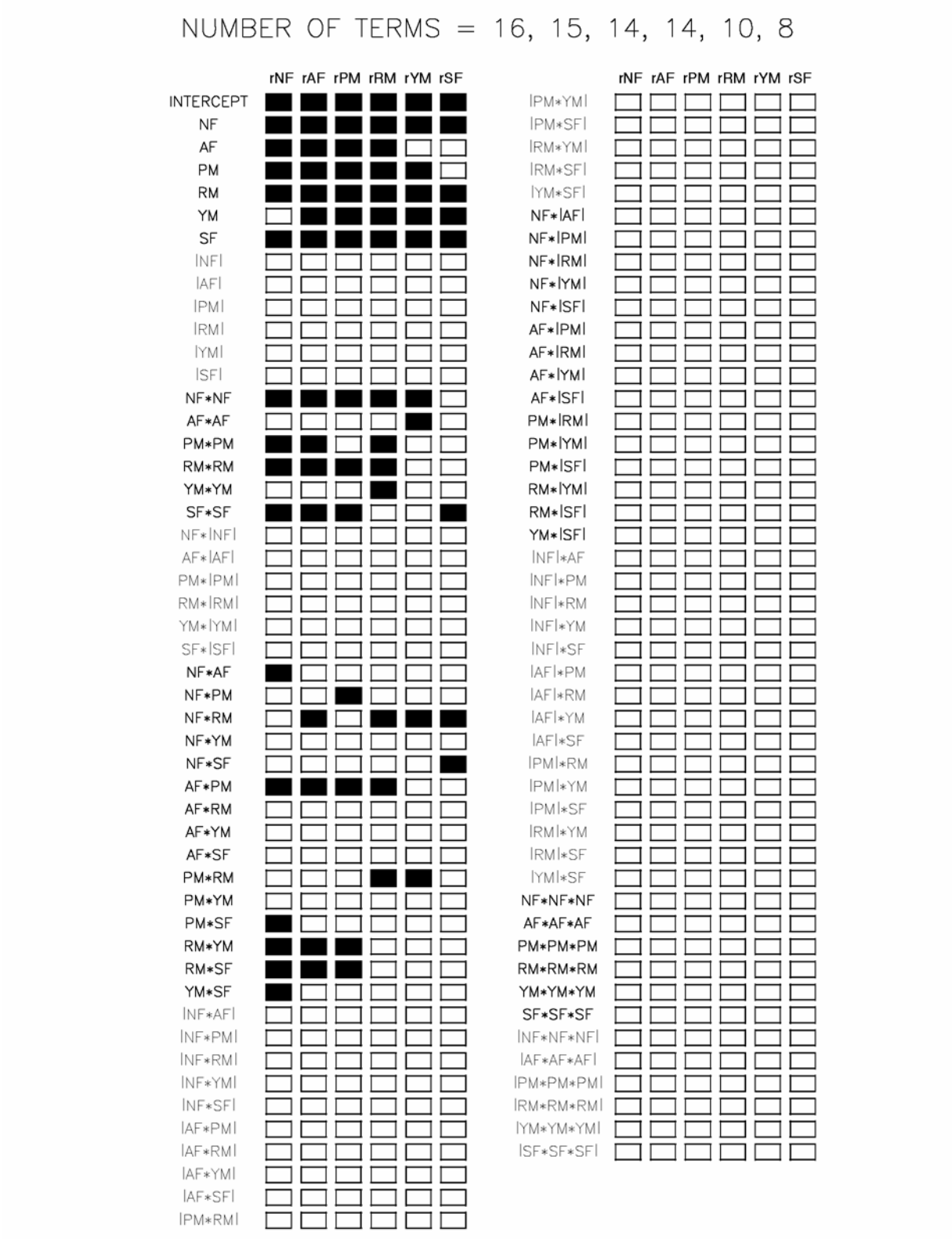

Figure 4d. Recommended math model used for analysis of Data Set 1 (BALFIT result).

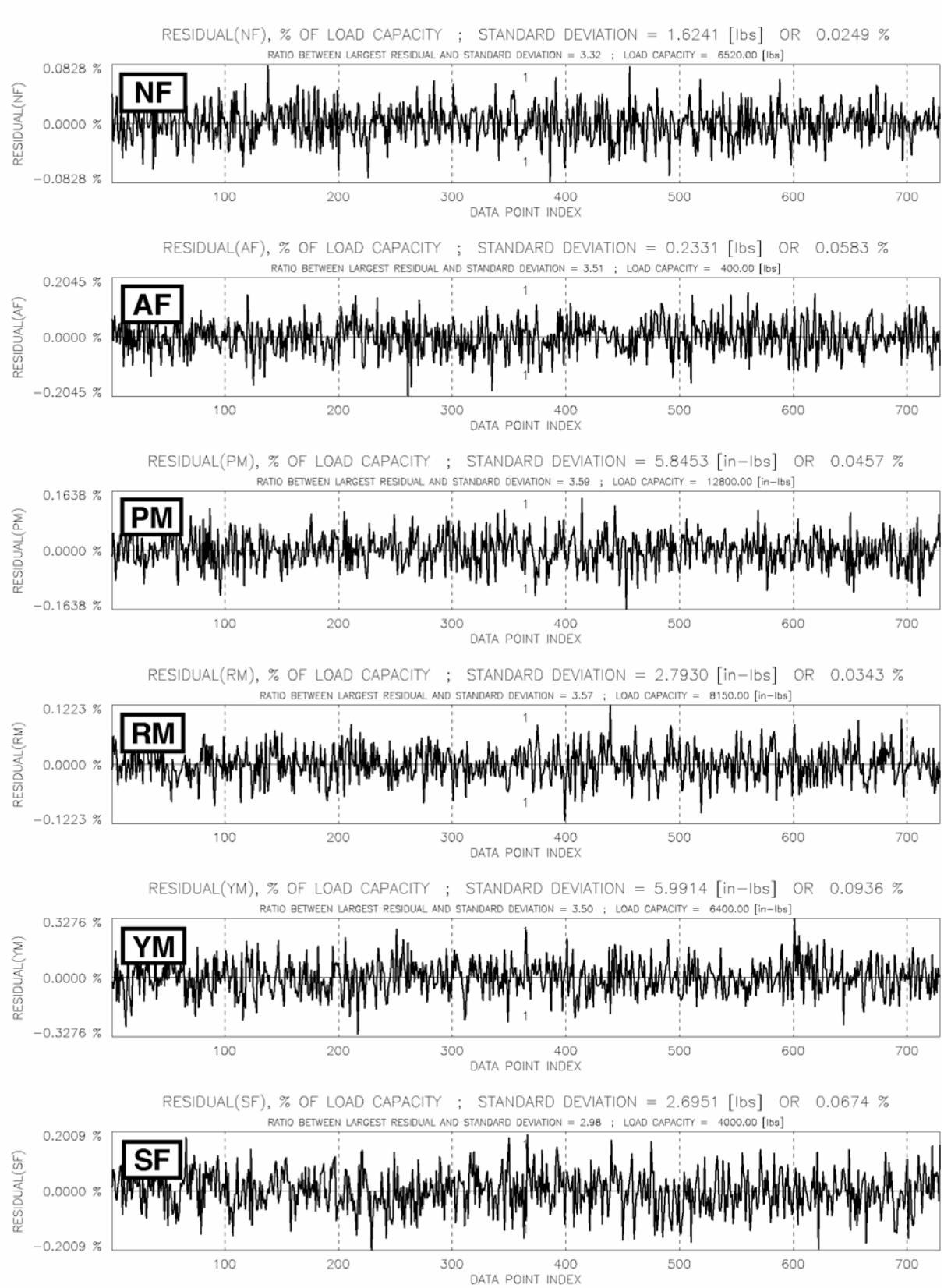

Figure 4e. Load residuals for recommended math model after analysis of Data Set 1 (BALFIT result).

Response 1 NF - Rnd **Transform:** None

Sequential Model Sum of Squares [Type I]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Mean vs Total	1197.5868	1	1197.5868			
Linear vs Mean	4.4546202E+008	6	74243669.	33350618.	< 0.0001	
2FI vs Linear	760.42210	15	50.694806	42.322377	< 0.0001	
Quadratic vs 2FI	649.95780	6	108.32630	385.65249	< 0.0001	Suggested
Cubic vs Quadra	13.170647	56	0.23519012	0.82563760	0.81334352	
Residual	183.73392	645	0.28485879			
Total	4.4546482E+008	729	611062.86			

"Sequential Model Sum of Squares [Type I]": Select the highest order polynomial where the additional terms are significant and the model is not aliased.

a)

Lack of Fit Tests

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Linear	1508.7176	398	3.7907476	12.460595	< 0.0001	
2FI	748.29546	383	1.9537740	6.4222652	< 0.0001	
Quadratic	98.337660	377	0.26084260	0.85741765	0.92501094	Suggested
Cubic	85.167013	321	0.26531780	0.87212810	0.89004569	
Pure Error	98.566903	324	0.30421884			

"Lack of Fit Tests": Want the selected model to have insignificant lack-of-fit.

b)

Model Summary Statistics

Source	Std. Dev.	R-Squared	Adjusted R-Squared	Predicted R-Squared	PRESS	
Linear	1.4920307	0.99999639	0.99999636	0.99999629	1652.1473	
2FI	1.0944520	0.99999810	0.99999804	0.99999791	931.27189	
Quadratic	0.52999147	0.99999956	0.99999954	0.99999952	213.13436	Suggested
Cubic	0.53372164	0.99999959	0.99999953	0.99999946	240.15686	

"Model Summary Statistics": Focus on the model maximizing the "Adjusted R-Squared" and the "Predicted R-Squared".

c)

Figure 5. Design Expert® NF Data Set 1. a) Sequential Model Sum of Squares report; b) Lack of Fit Tests report; c) Model Summary Statistics report.

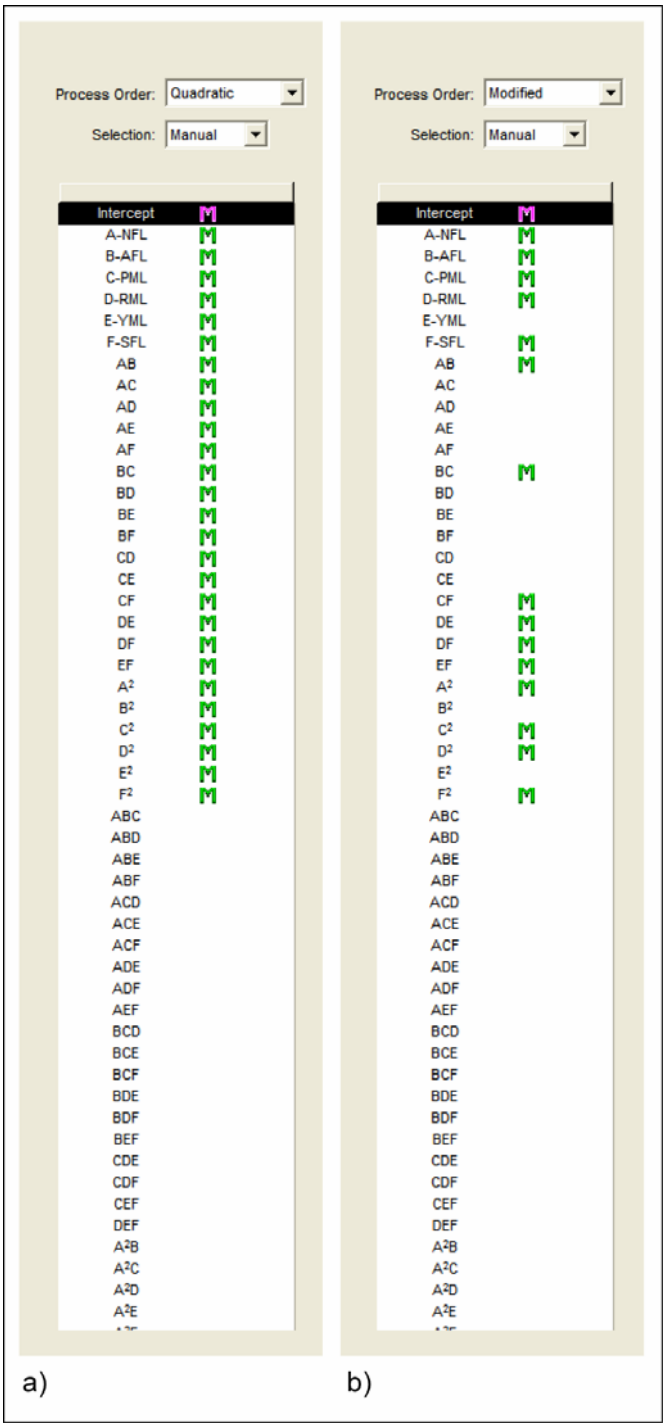

Figure 6. Design Expert® NF Models for Data Set 1. a) "Permitted Model"; b) "Recommended Model."

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	4.455E+008	15	2.970E+007	1.064E+008	< 0.0001
A-NFL	*1.049E+008*	*1*	*1.049E+008*	*3.760E+008*	*< 0.0001*
B-AFL	*377.33*	*1*	*377.33*	*1351.96*	*< 0.0001*
C-PML	*931.56*	*1*	*931.56*	*3337.78*	*< 0.0001*
D-RML	*407.95*	*1*	*407.95*	*1461.69*	*< 0.0001*
F-SFL	*5.65*	*1*	*5.65*	*20.23*	*< 0.0001*
AB	*4.14*	*1*	*4.14*	*14.85*	*0.0001*
BC	*8.76*	*1*	*8.76*	*31.39*	*< 0.0001*
CF	*12.91*	*1*	*12.91*	*46.26*	*< 0.0001*
DE	*4.56*	*1*	*4.56*	*16.36*	*< 0.0001*
DF	*720.84*	*1*	*720.84*	*2582.78*	*< 0.0001*
EF	*9.40*	*1*	*9.40*	*33.70*	*< 0.0001*
A^2	*328.58*	*1*	*328.58*	*1177.29*	*< 0.0001*
C^2	*25.63*	*1*	*25.63*	*91.84*	*< 0.0001*
D^2	*77.50*	*1*	*77.50*	*277.67*	*< 0.0001*
F^2	*46.63*	*1*	*46.63*	*167.06*	*< 0.0001*
Residual	199.00	713	0.28		
Lack of Fit	*100.43*	*389*	*0.26*	*0.85*	*0.9392*
Pure Error	*98.57*	*324*	*0.30*		
Cor Total	4.455E+008	728			

Std. Dev.	0.53	R-Squared	1.0000
Mean	-1.28	Adj R-Squared	1.0000
C.V. %	41.22	Pred R-Squared	1.0000
PRESS	207.81	Adeq Precision	55469.722

Figure 7a. Design Expert® NF ANOVA for Data Set 1.

Factor	Coefficient Estimate	df	Standard Error	95% CI Low	95% CI High	VIF
Intercept	-0.95	1	0.032	-1.01	-0.88	
A-NFL	2096.78	1	0.11	2096.57	2097.00	4.25
B-AFL	-1.00	1	0.027	-1.05	-0.94	1.27
C-PML	5.10	1	0.088	4.93	5.27	2.15
D-RML	-1.43	1	0.037	-1.50	-1.36	1.00
F-SFL	-0.23	1	0.050	-0.33	-0.13	1.08
AB	0.42	1	0.11	0.21	0.63	4.25
BC	-0.49	1	0.088	-0.67	-0.32	2.14
CF	-1.84	1	0.27	-2.37	-1.31	1.50
DE	-0.45	1	0.11	-0.68	-0.23	1.00
DF	-7.27	1	0.14	-7.55	-6.99	1.00
EF	0.62	1	0.11	0.41	0.83	1.09
A^2	2.16	1	0.063	2.03	2.28	1.15
C^2	0.80	1	0.084	0.64	0.97	1.04
D^2	-0.89	1	0.053	-0.99	-0.78	1.28
F^2	-0.83	1	0.064	-0.96	-0.71	1.22

Figure 7b. Design Expert® NF Regression Coefficients for Data Set 1.

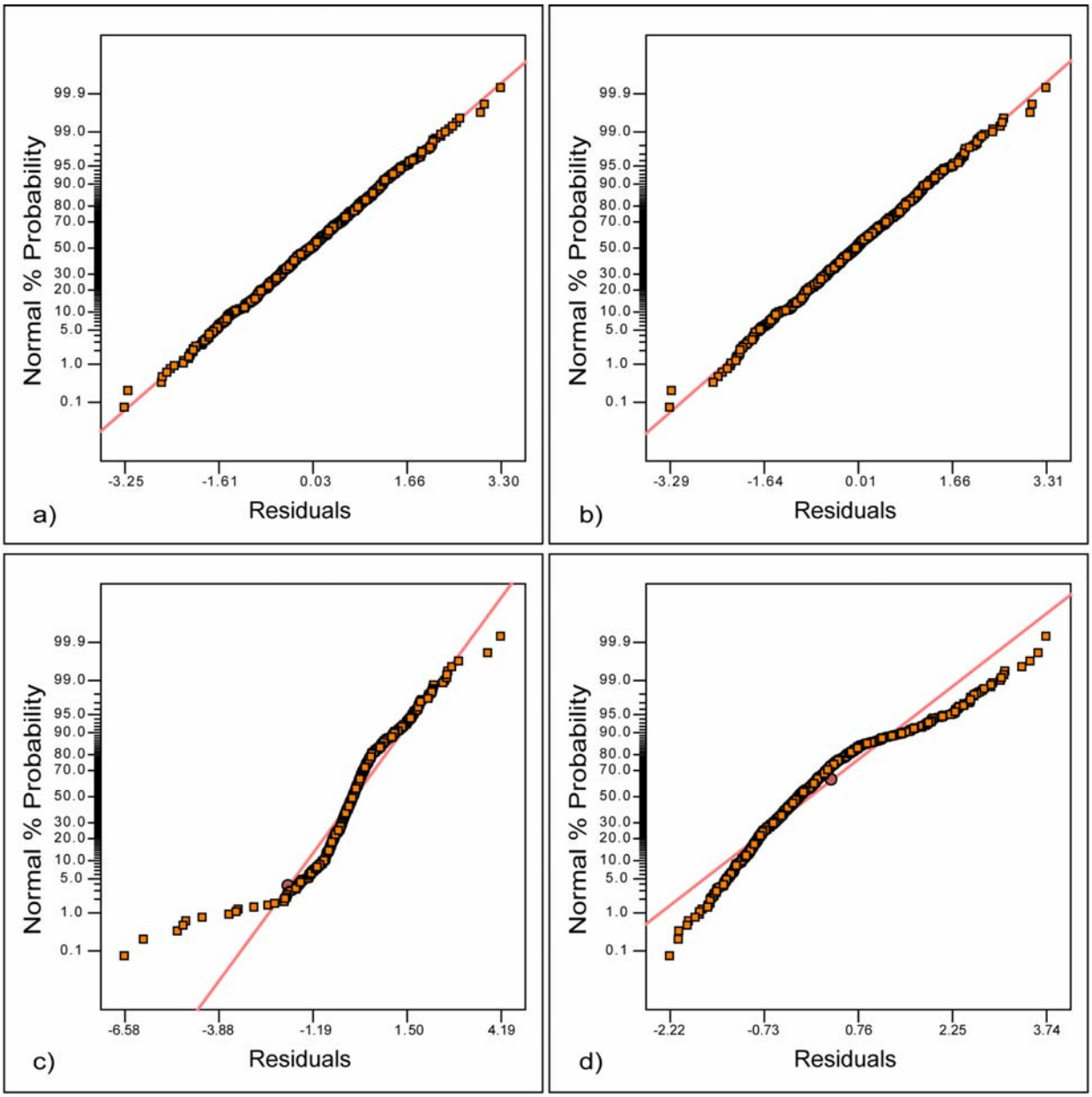

Figure 8. Design Expert® NF Normal Probability Plot of Residuals for Data Set 1. a) 17-Term Recommended Model; b) Full 28-Term Quadratic Model; c) 6-Term Linear Model; d) 22-Term Factor Interaction Model.

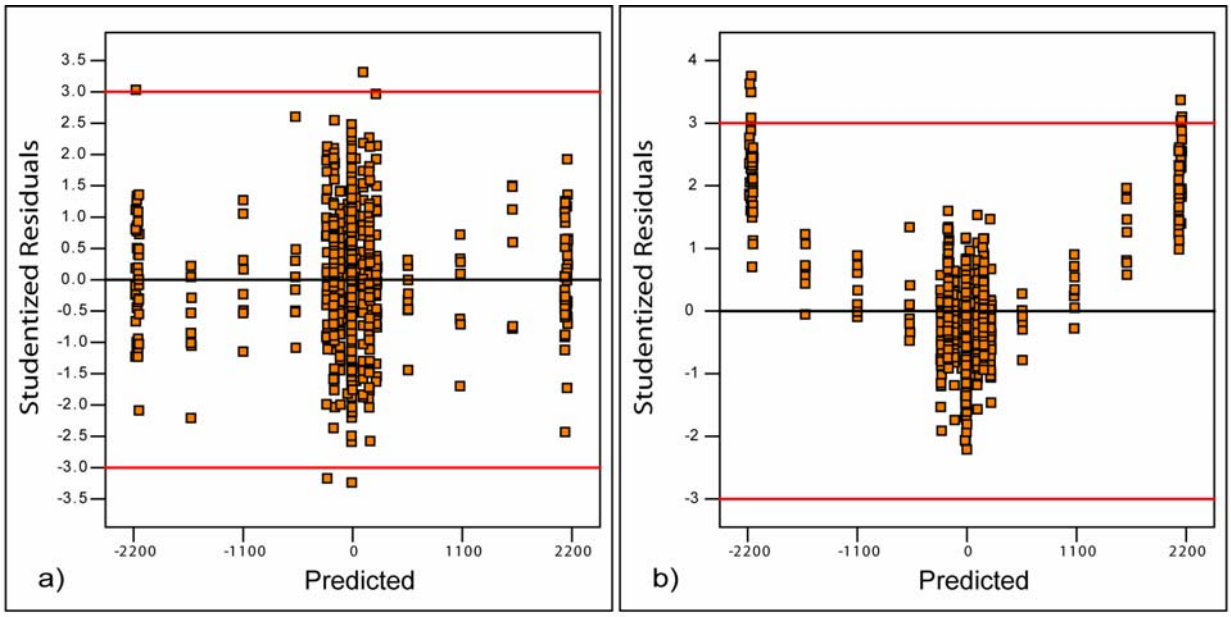

Figure 9. Design Expert® NF Plot of Residuals vs. Predicted Responses for Data Set 1. a) Recommended Model, b) Factor Interaction Model.

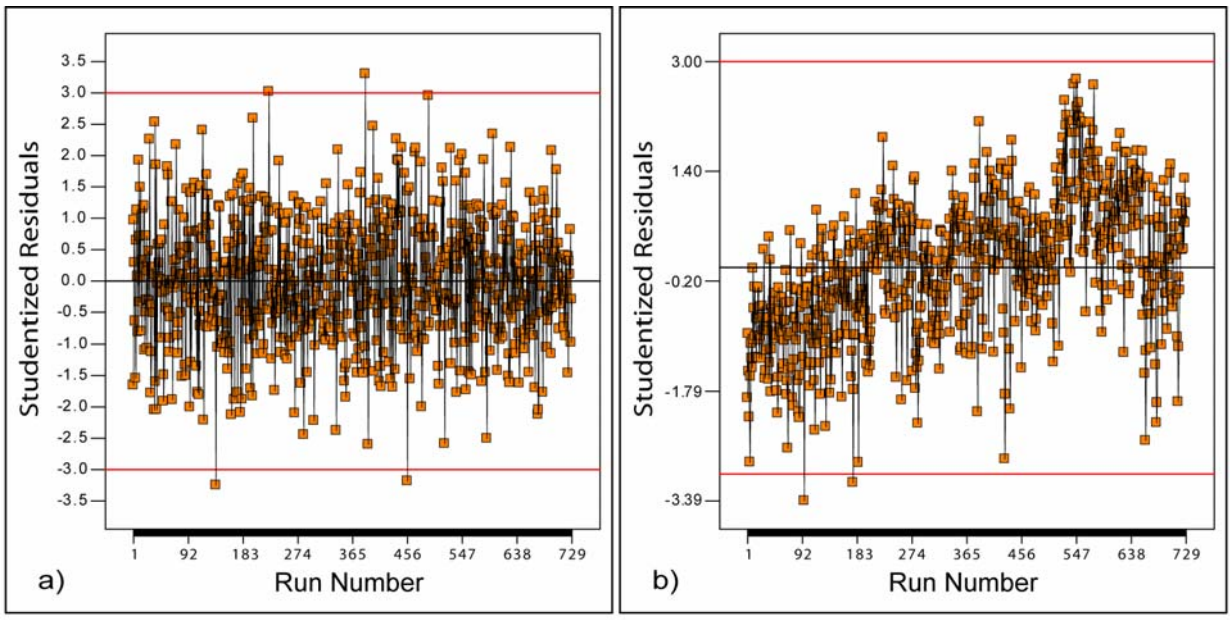

Figure 10. Design Expert® NF Plot of Residuals vs. Run Number (Time) for Data Set 1. a) Random error only, b) Random-plus-Systematic error.

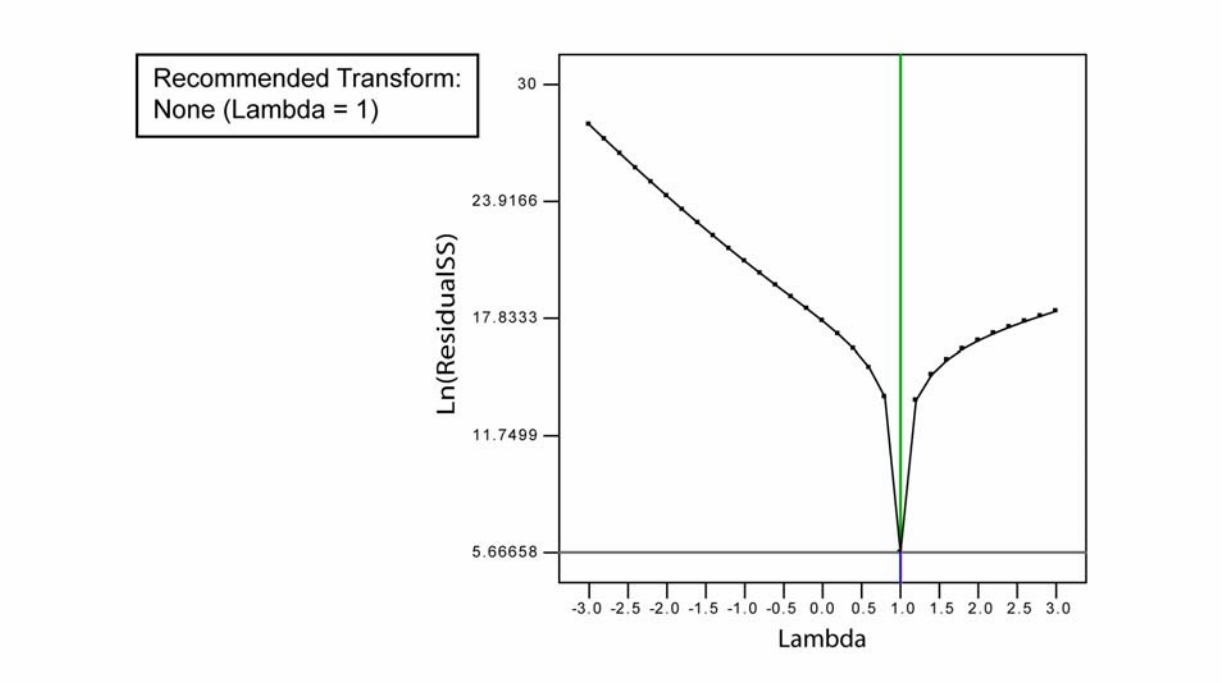

Figure 11. Design Expert® Representative Box-Cox Transformation Plot.

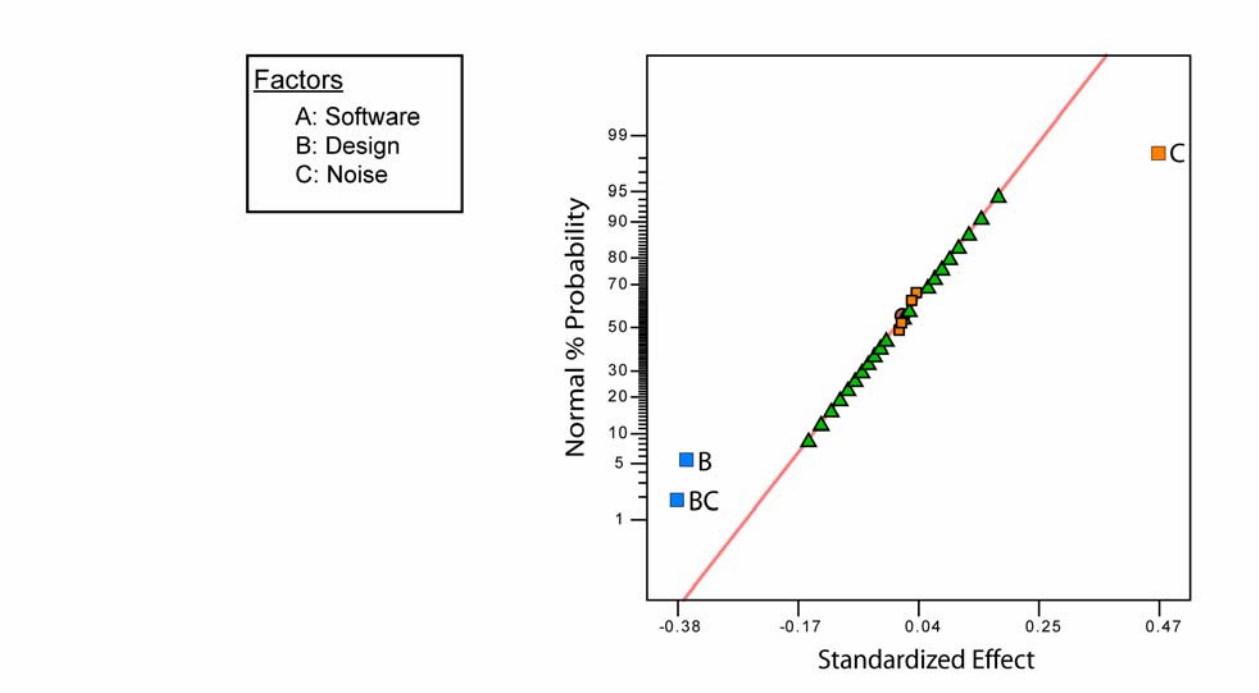

Figure 12. Normal Probability Plot: Standard Deviation of Calibration Residuals.

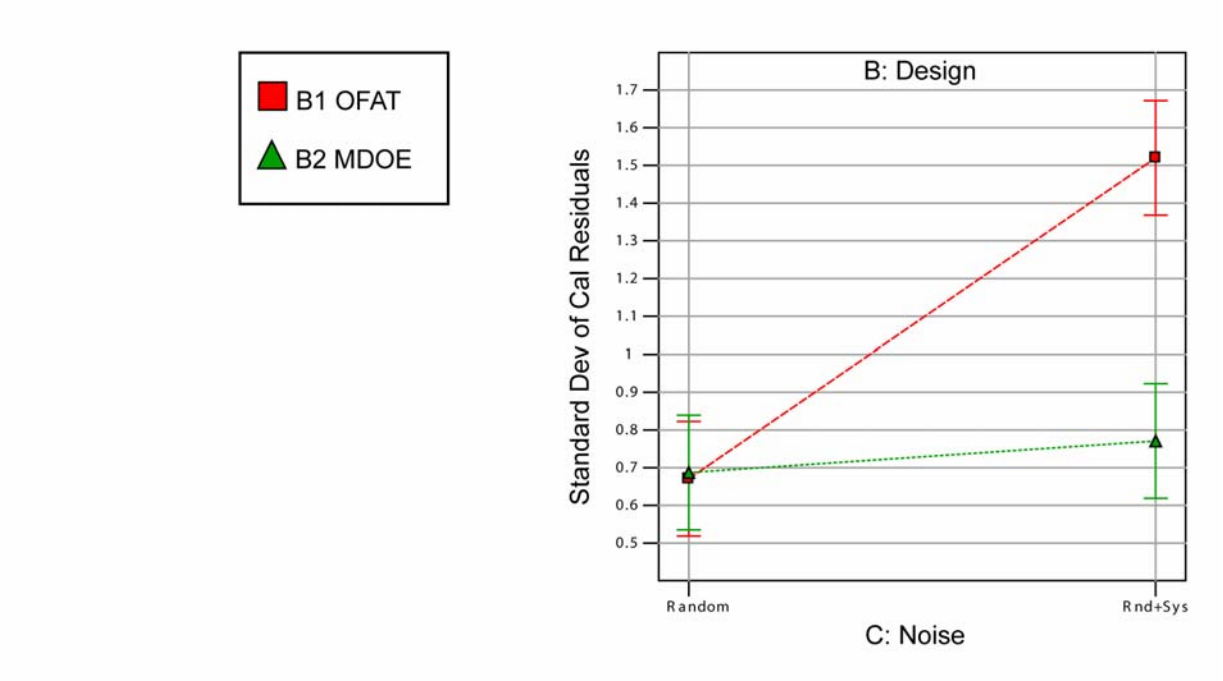

Figure 13. Interaction Graph: Standard Deviation of Calibration Residuals.

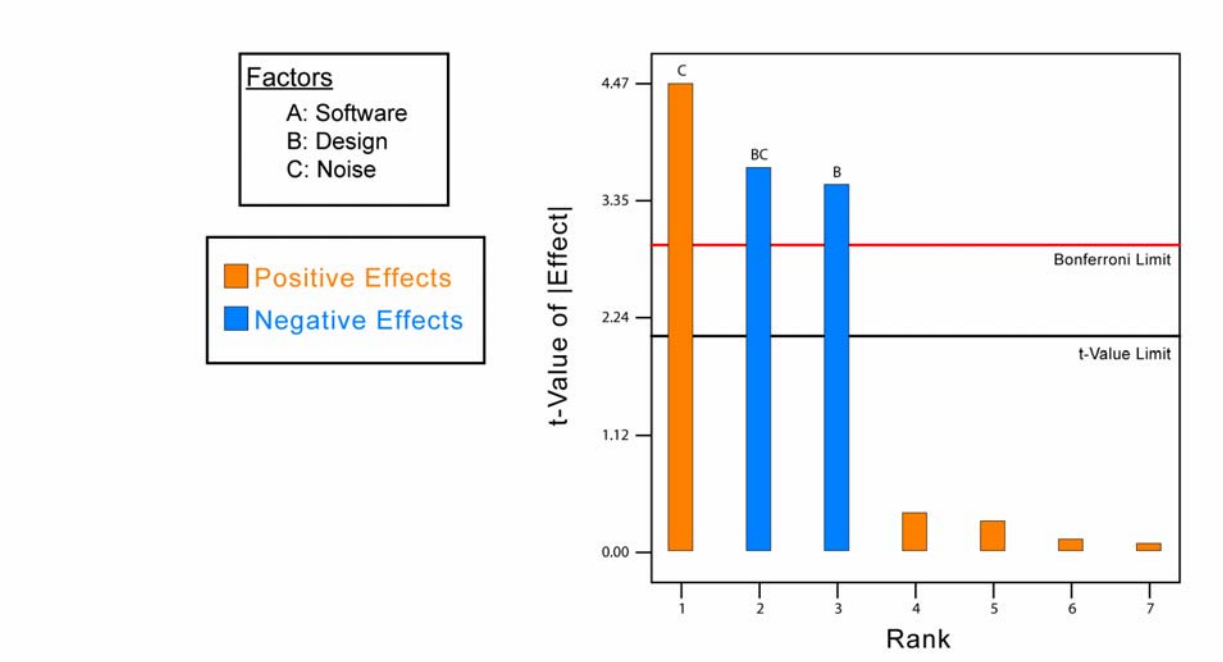

Figure 14. Pareto Chart: Standard Deviation of Calibration Residuals.

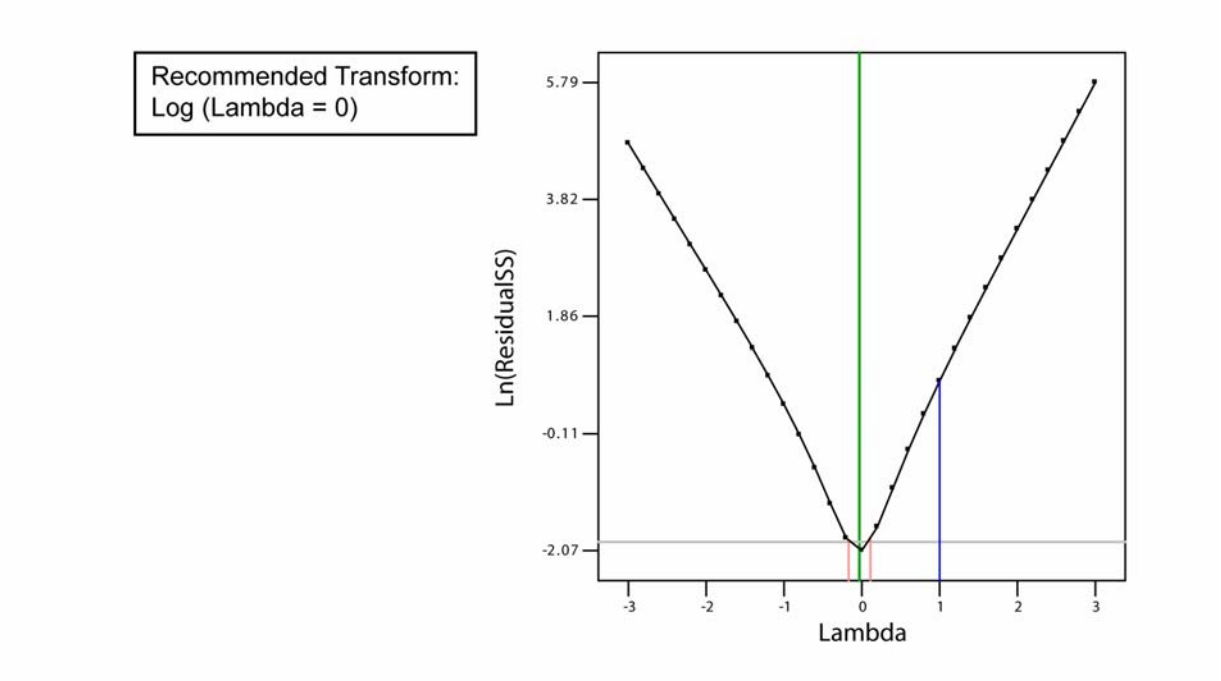

Figure 15. Box-Cox Transformation Plot: Standard Deviation of Calibration Residuals.

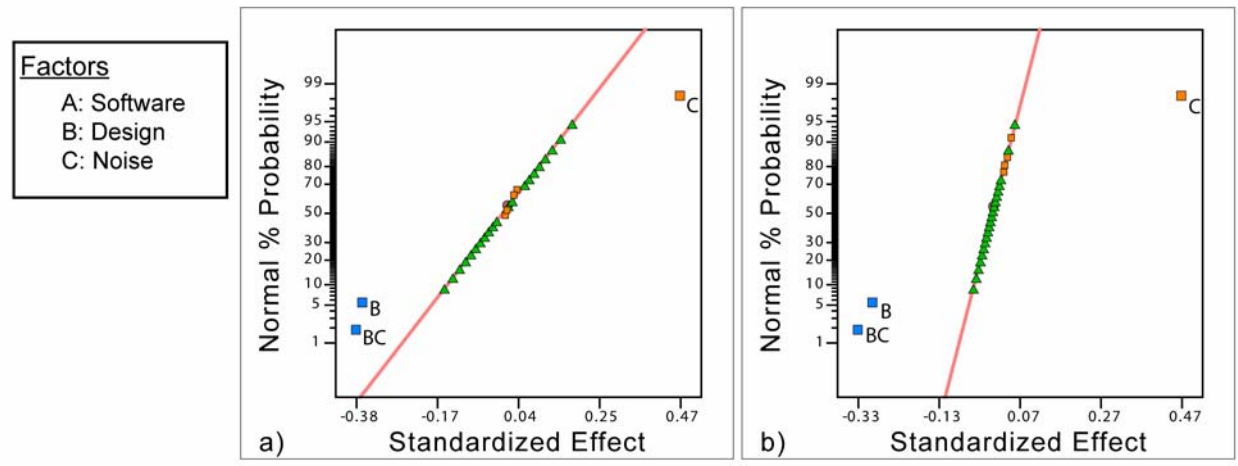

Figure 16. Normal Probability Plot: Standard Deviation of Calibration Residuals. a) Untransformed, b) logarithmic transformation.

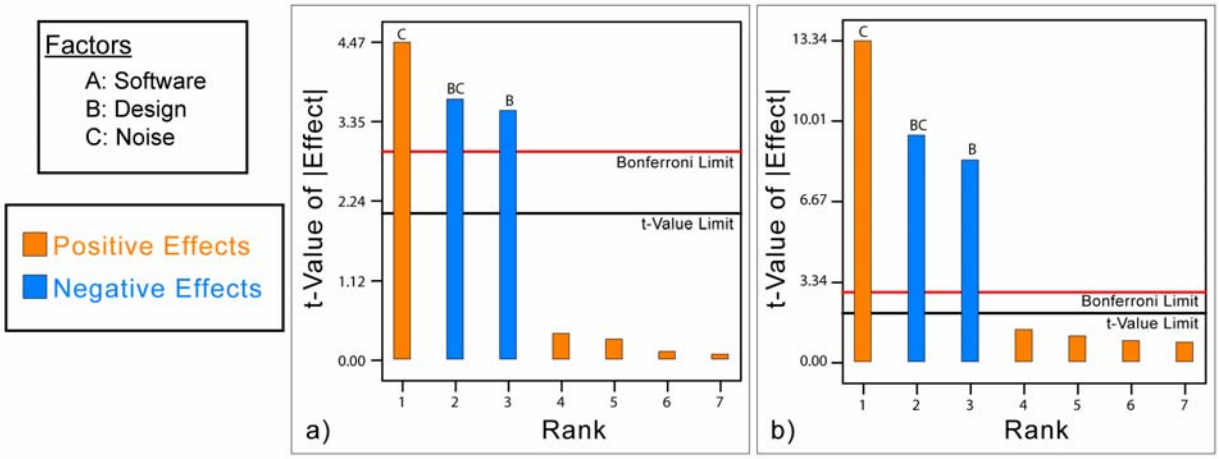

Figure 17. Pareto Chart: Standard Deviation of Calibration Residuals. a) Untransformed, b) logarithmic transformation.

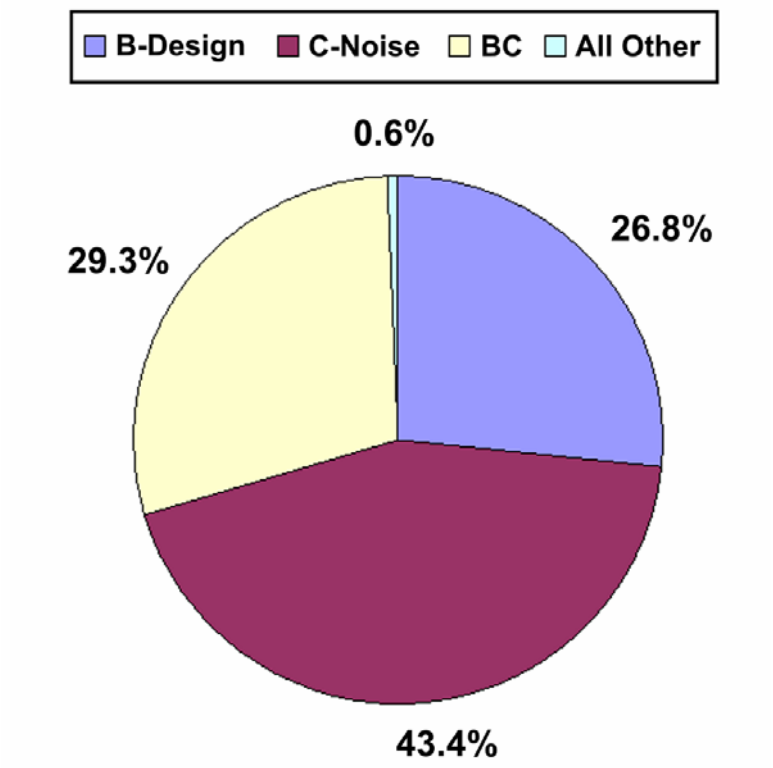

Figure 18. Partition of Explained Sum of Squares: Standard Deviation of Calibration Residuals.

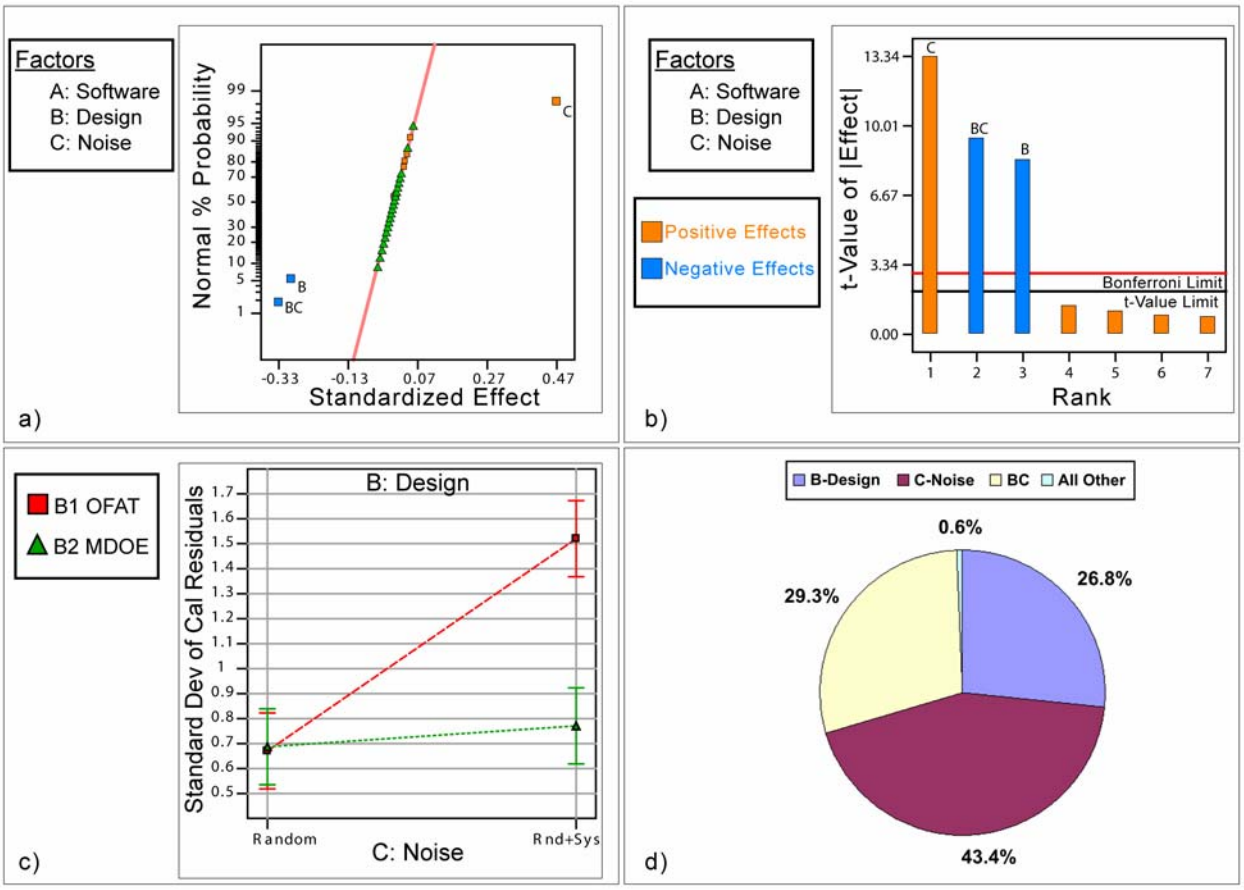

Figure 19. Factor Effects for Standard Deviation of Calibration Residuals. a) Normal Probability Plot, b) Pareto Chart, c) Interaction Graph, d) Partition of Explained Sum of Squares.

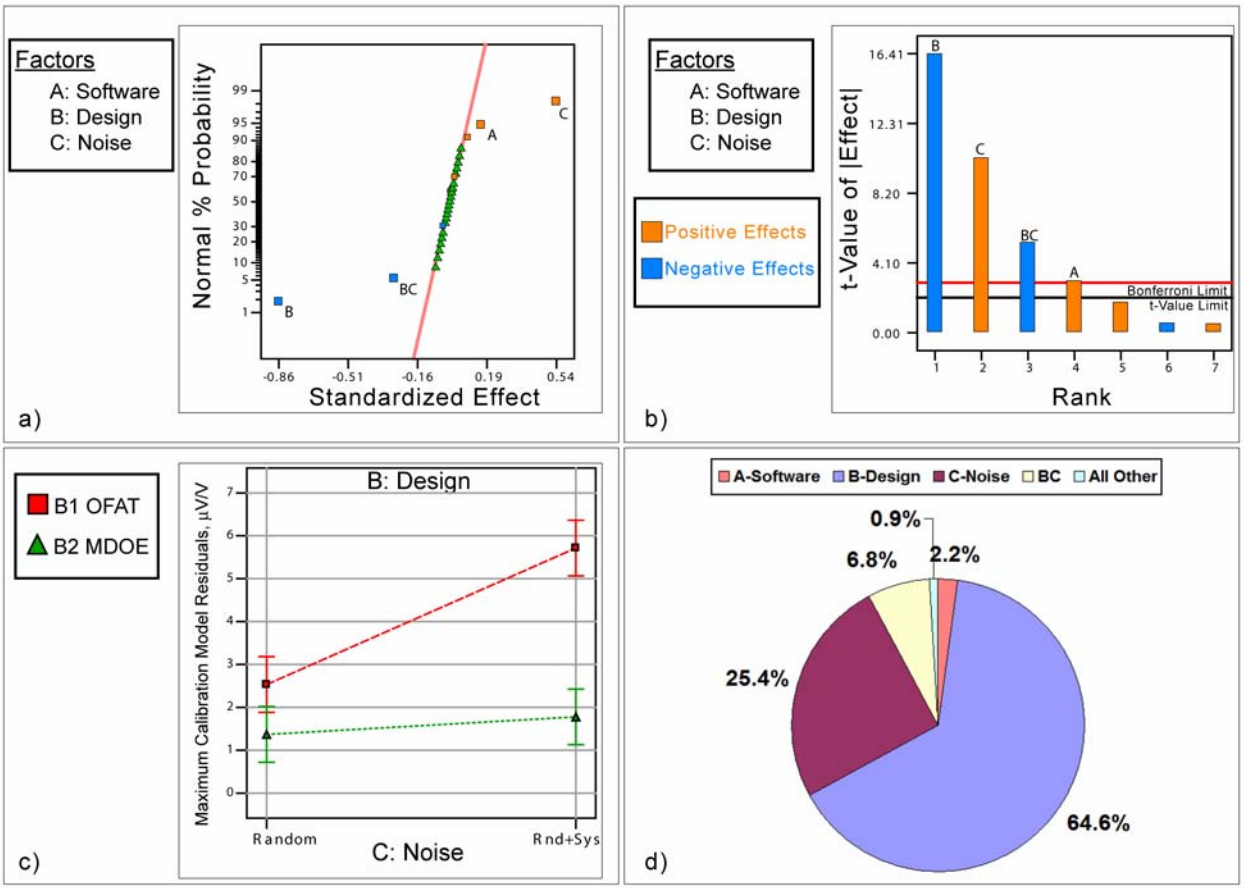

Figure 20. a) Normal Probability Plot: Maximum Model Residual; b) Pareto Chart: Maximum Calibration Residuals; c) Interaction Graph: Maximum Calibration Residuals; d) Partition of Explained Sum of Squares: Maximum Calibration Residuals.

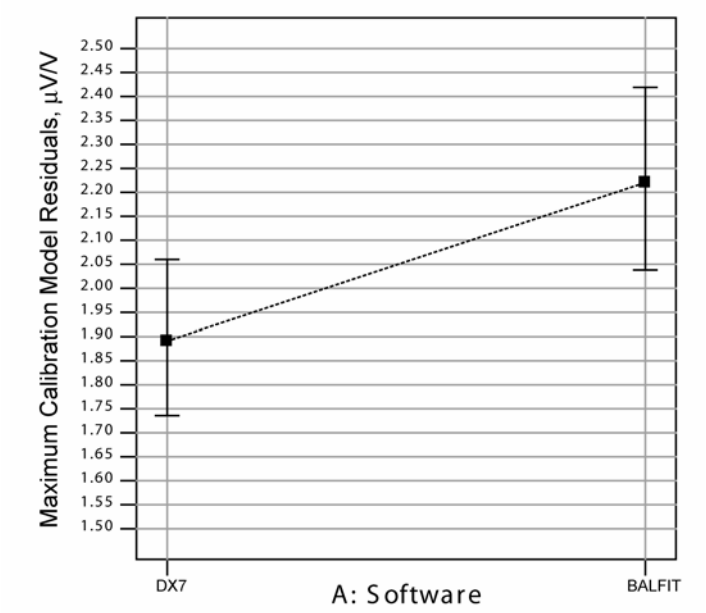

Figure 21. Software Main Effect: Maximum Calibration Residuals. Least Significant Difference Bars Just Overlap. Cannot resolve software effect with high confidence.

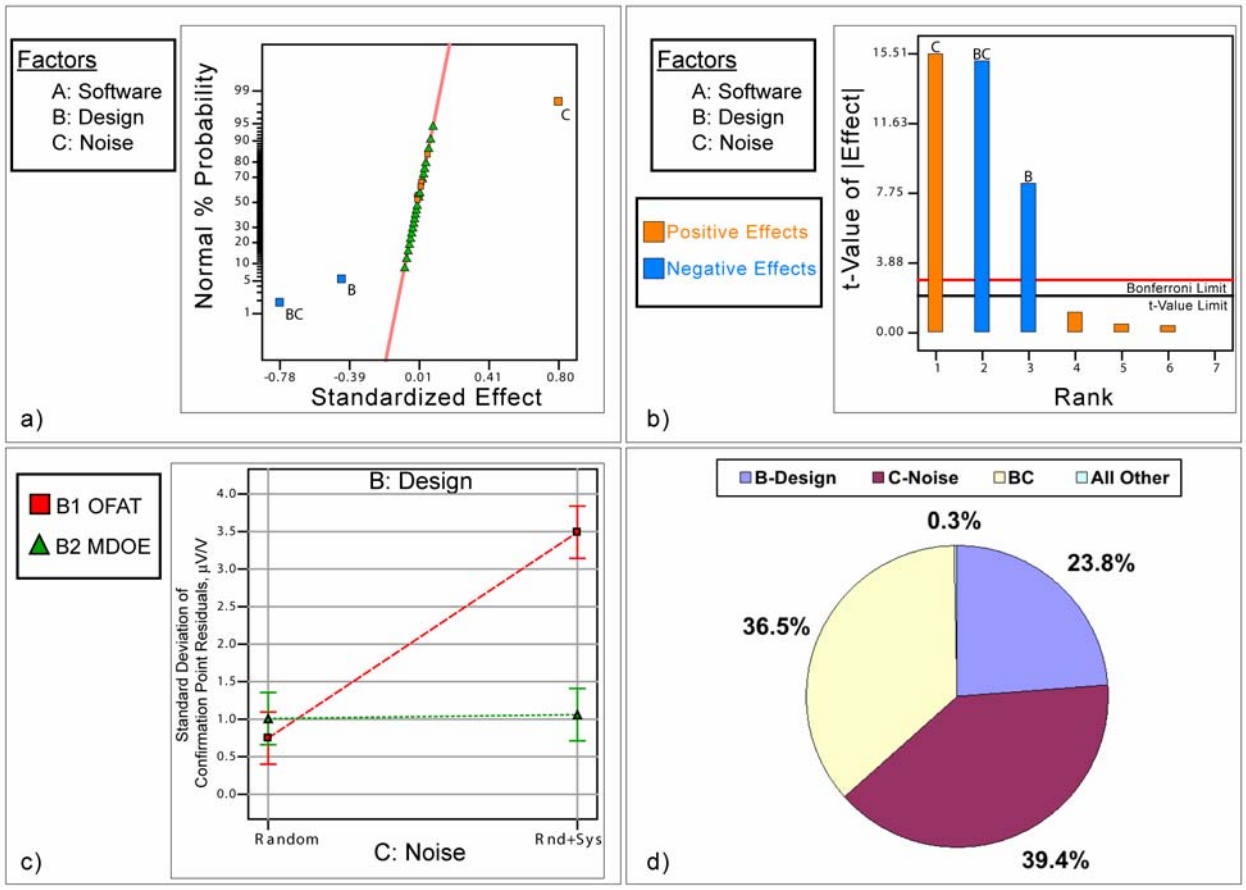

Figure 22. a) Normal Probability Plot: Standard Deviation of Confirmation Point Residuals; b) Pareto Chart: Standard Deviation of Confirmation Point Residuals; c) Interaction Graph: Standard Deviation of Confirmation Point Residuals; d) Partition of Explained Sum of Squares: Standard Deviation of Confirmation Point Residuals.

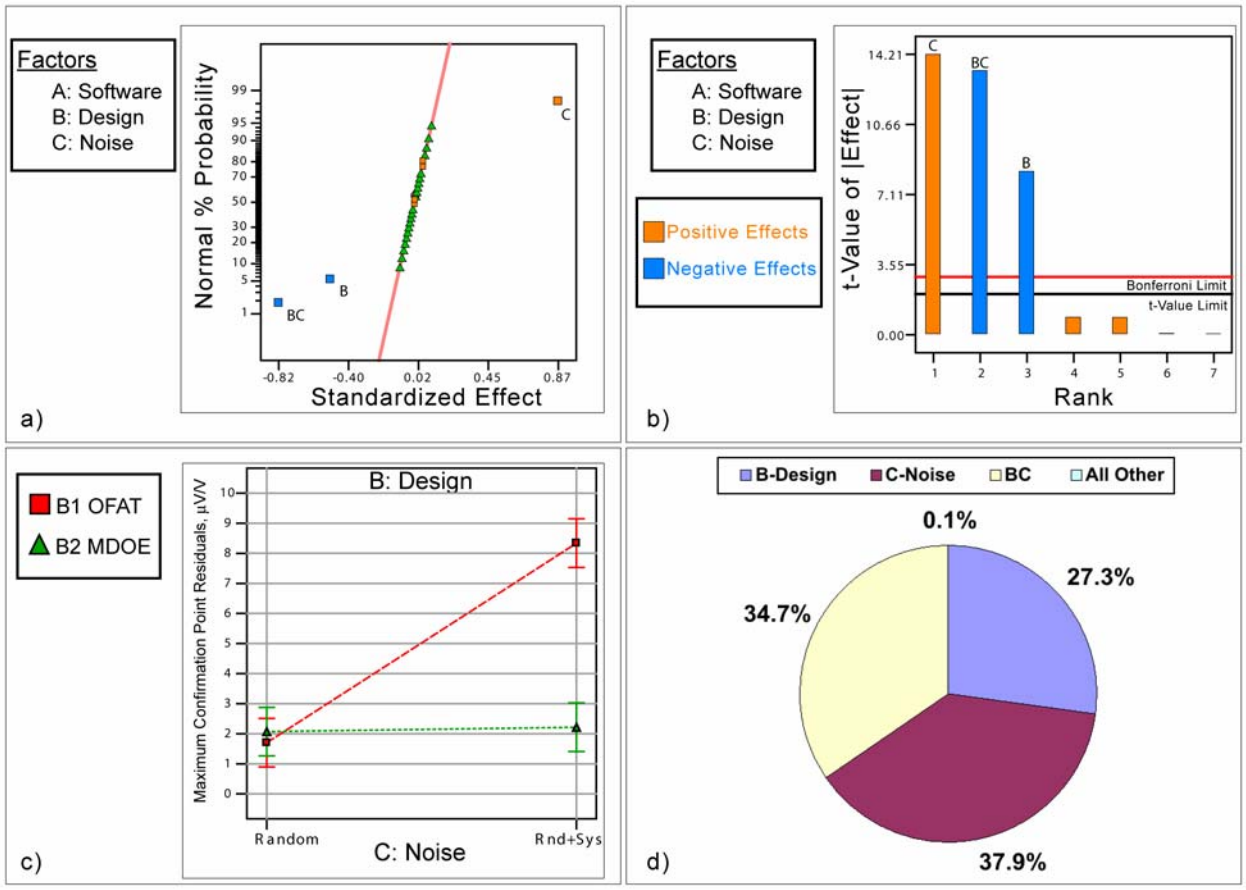

Figure 23. a) Normal Probability Plot: Maximum Confirmation Point Residuals; b) Pareto Chart: Maximum Confirmation Point Residuals; c) Interaction Graph: Maximum Confirmation Point Residuals; d) Partition of Explained Sum of Squares: Maximum Confirmation Point Residuals.

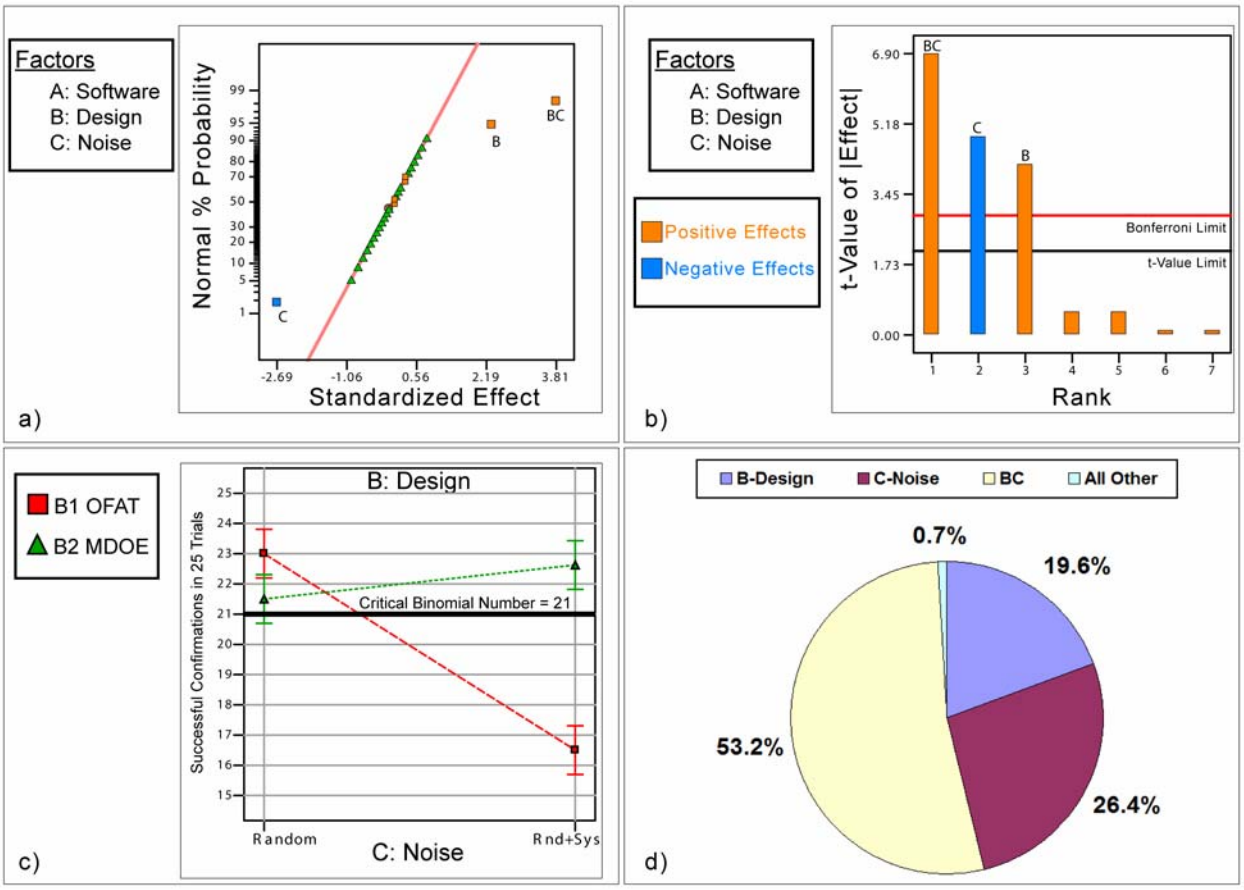

Figure 24. a) Normal Probability Plot: Successful Confirmations; b) Pareto Chart: Successful Confirmations; c) Interaction Graph: Successful Confirmations; d) Partition of Explained Sum of Squares: Successful Confirmations.

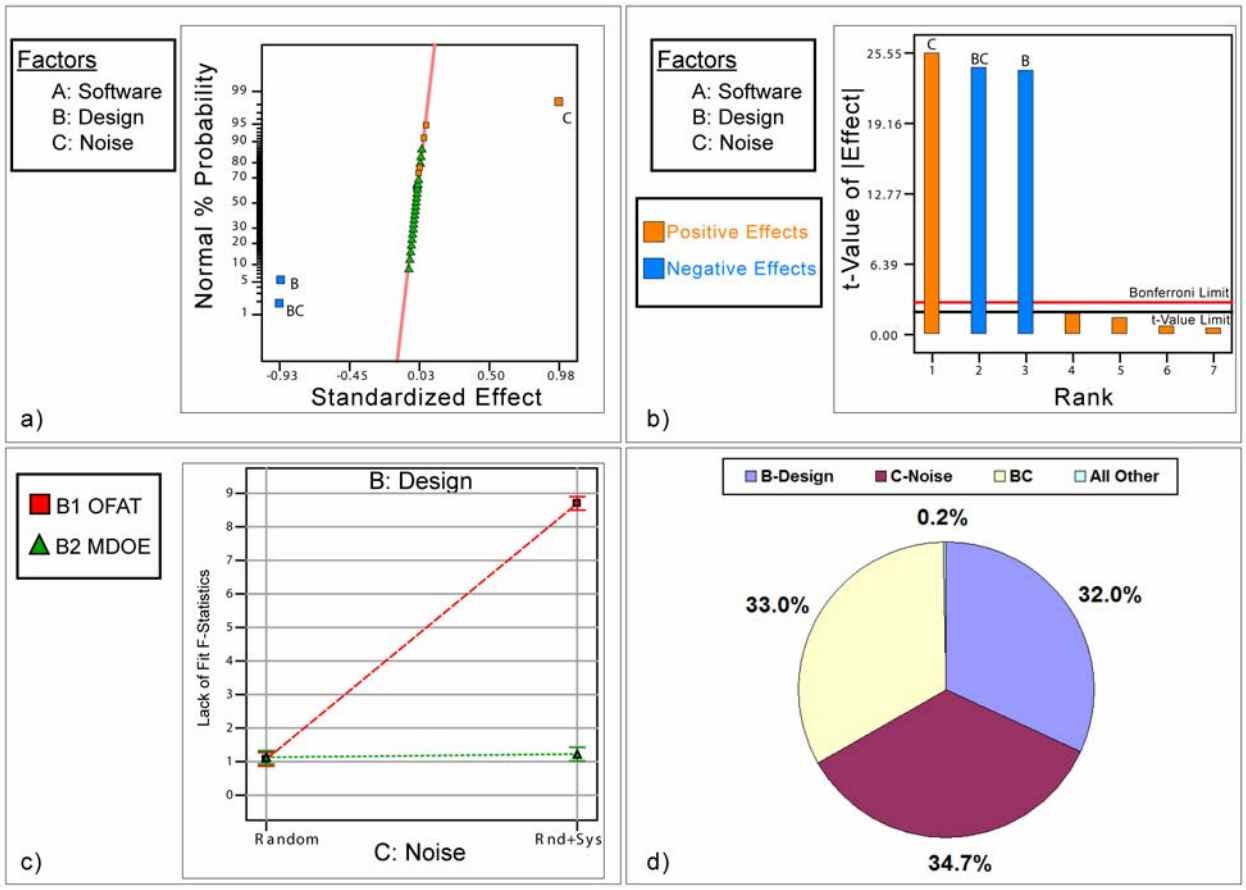

Figure 25. a) Normal Probability Plot: Lack of Fit F-Statistic Effects; b) Pareto Chart: Lack of Fit F-Statistic Effects; c) Interaction Graph: Lack of Fit F-Statistic Effects; d) Partition of Explained Sum of Squares: Lack of Fit F-Statistic Effects.

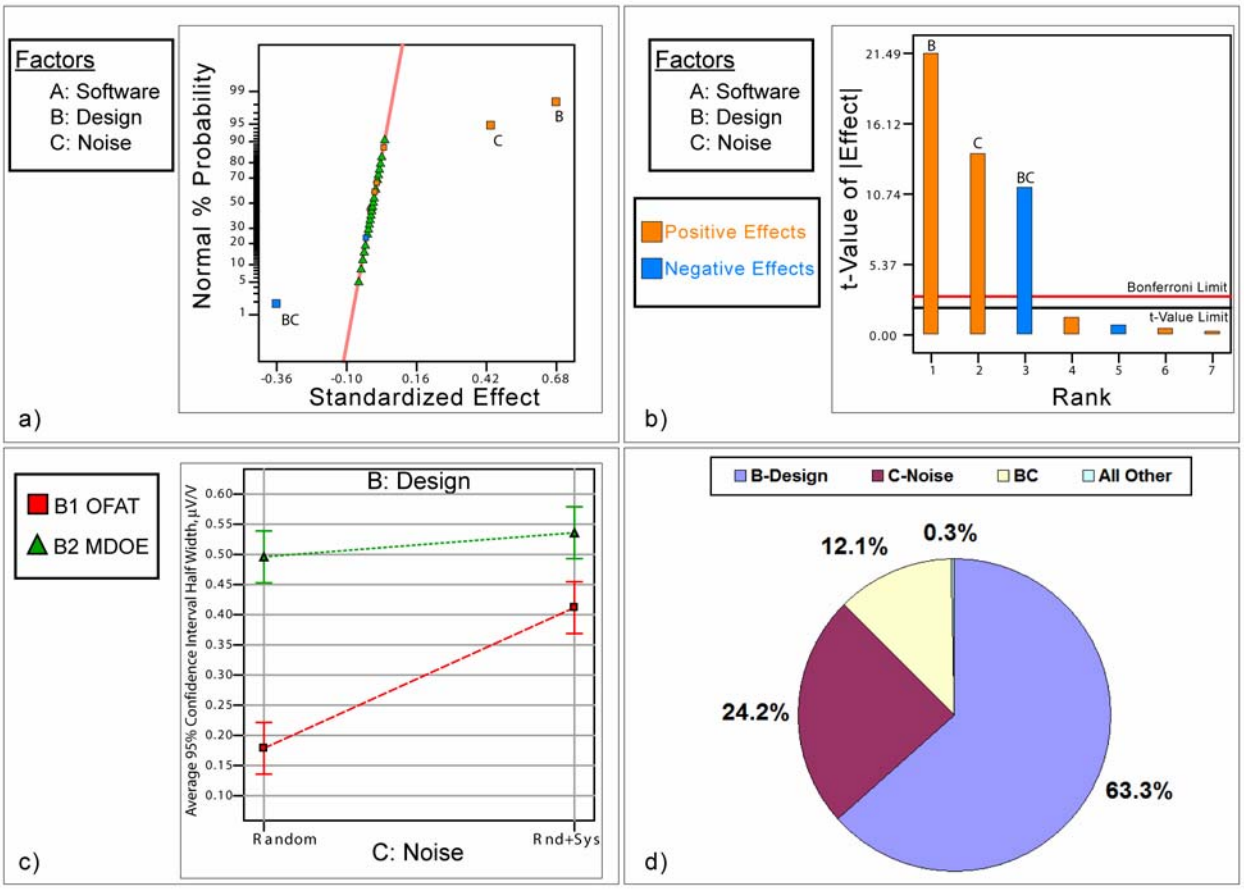

Figure 26. a) Normal Probability Plot: Average 95% Confidence Interval Half Width for Model Predictions; b) Pareto Chart: Average 95% Confidence Interval Half Width for Model Predictions; c) Interaction Graph: Average 95% Confidence Interval Half Width for Model Predictions; d) Partition of Explained Sum of Squares: Average 95% Confidence Interval Half Width for Model Predictions.

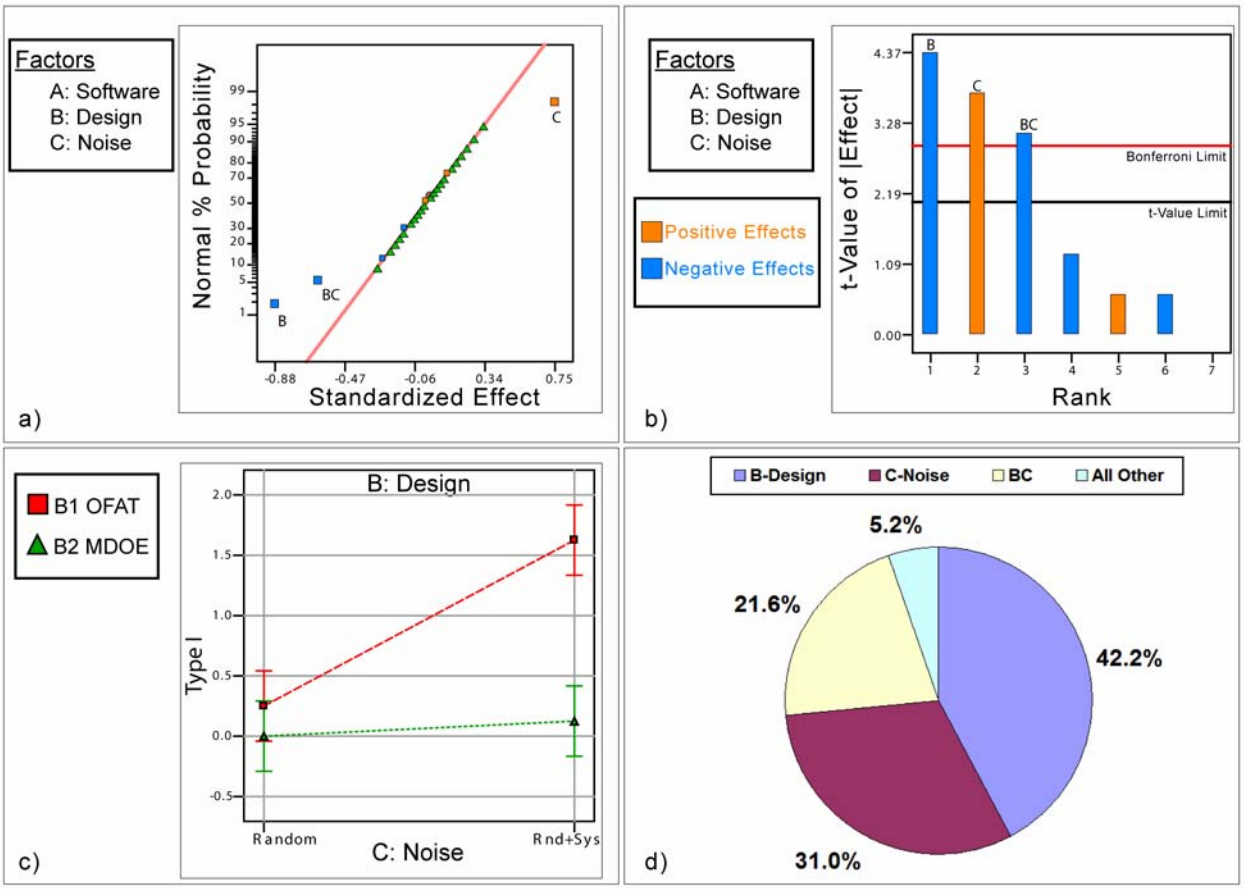

Figure 27. a) Normal Probability Plot: Number of Type I Inference Errors (Including Terms That Do Not Belong in Model); b) Pareto Chart: Number of Type I Inference Errors (Including Terms That Do Not Belong in Model); c) Interaction Graph: Number of Type I Inference Errors (Including Terms That Do Not Belong in Model); d) Partition of Explained Sum of Squares: Number of Type I Inference Errors (Including Terms That Do Not Belong in Model).

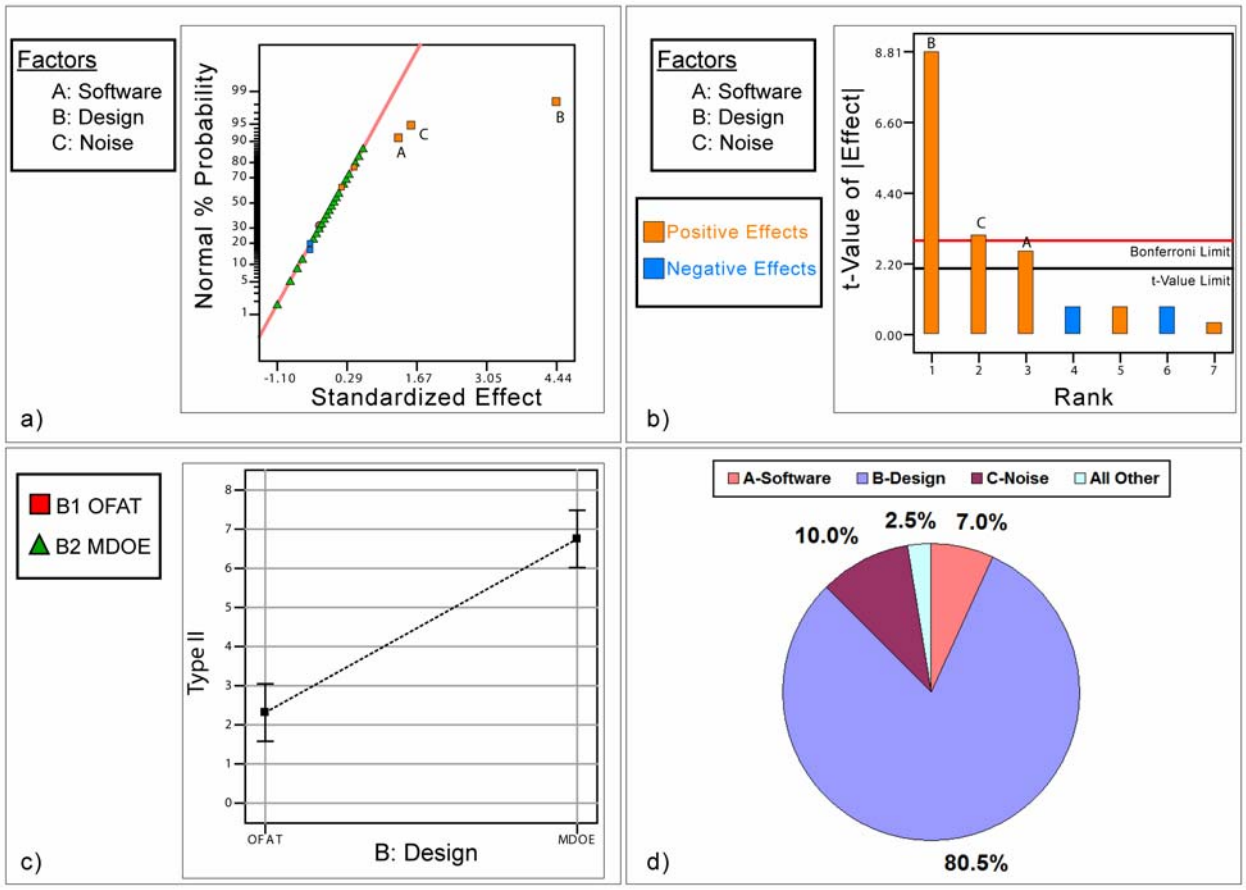

Figure 28. a) Normal Probability Plot: Number of Type II Inference Errors (Excluding Terms That Are in the "True" Model); b) Pareto Chart: Number of Type II Inference Errors (Excluding Terms That Are in the "True" Model); c) Main Design Effect: Number of Type II Inference Errors (Excluding Terms That Are in the "True" Model).MDOE Neglects Statistically Significant Terms if they are Too Small to be of Practical Interest; d) Partition of Explained Sum of Squares: Number of Type II Inference Errors (Excluding Terms That Are in the "True" Model).

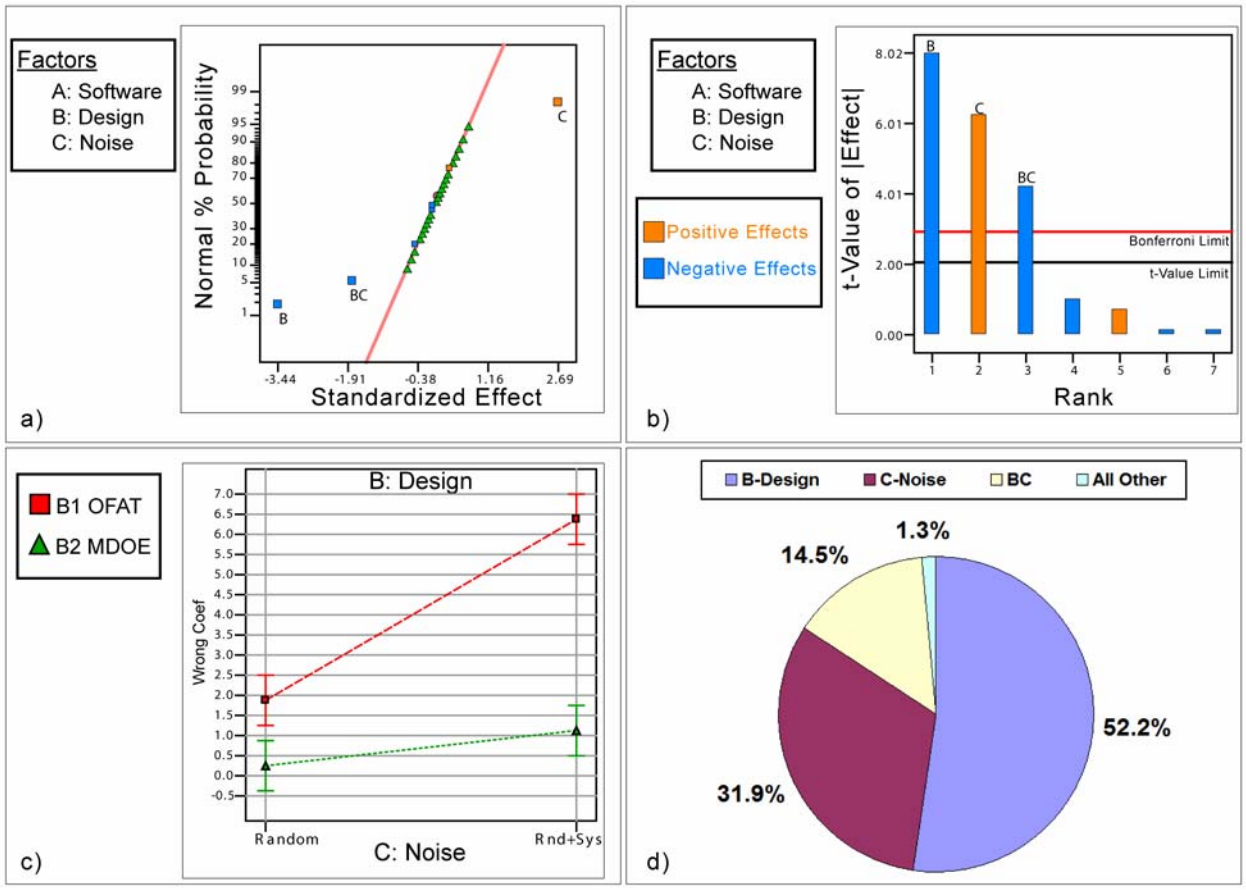

Figure 29. a) Normal Probability Plot: Number of Erroneously Estimated Coefficients; b) Pareto Chart: Number of Erroneously Estimated Coefficients; c) Main Design Effect: Number of Erroneously Estimated Coefficients; d) Partition of Explained Sum of Squares: Number of Erroneously Estimated Coefficients.

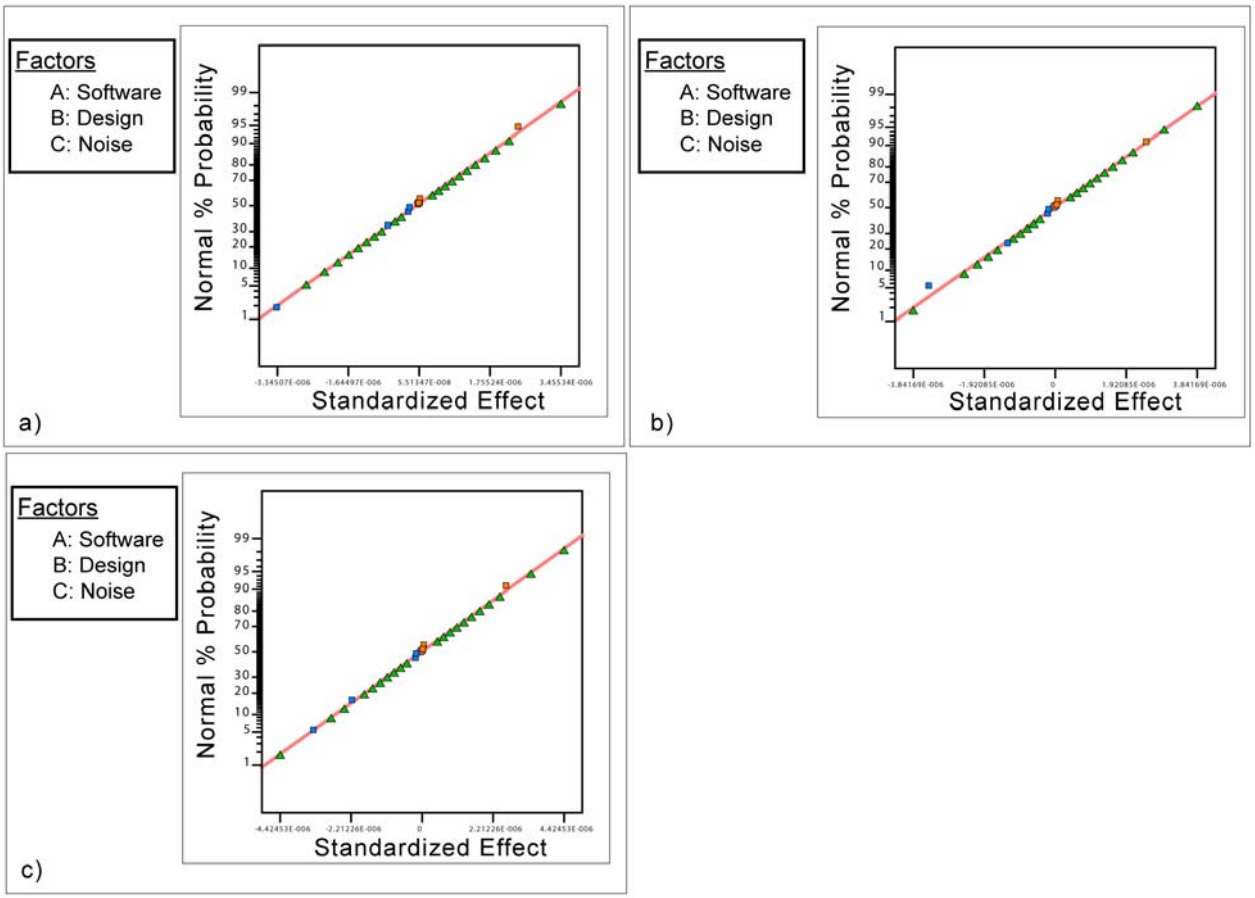

Figure 30. a) Normal Probability Plot: Ordinary R-Squared Effects. No Effects Are Significant; b) Normal Probability Plot: Adjusted R-Squared Effects. No Effects Are Significant; c) Normal Probability Plot: Predicted R-Squared Effects. No Effects Are Significant.

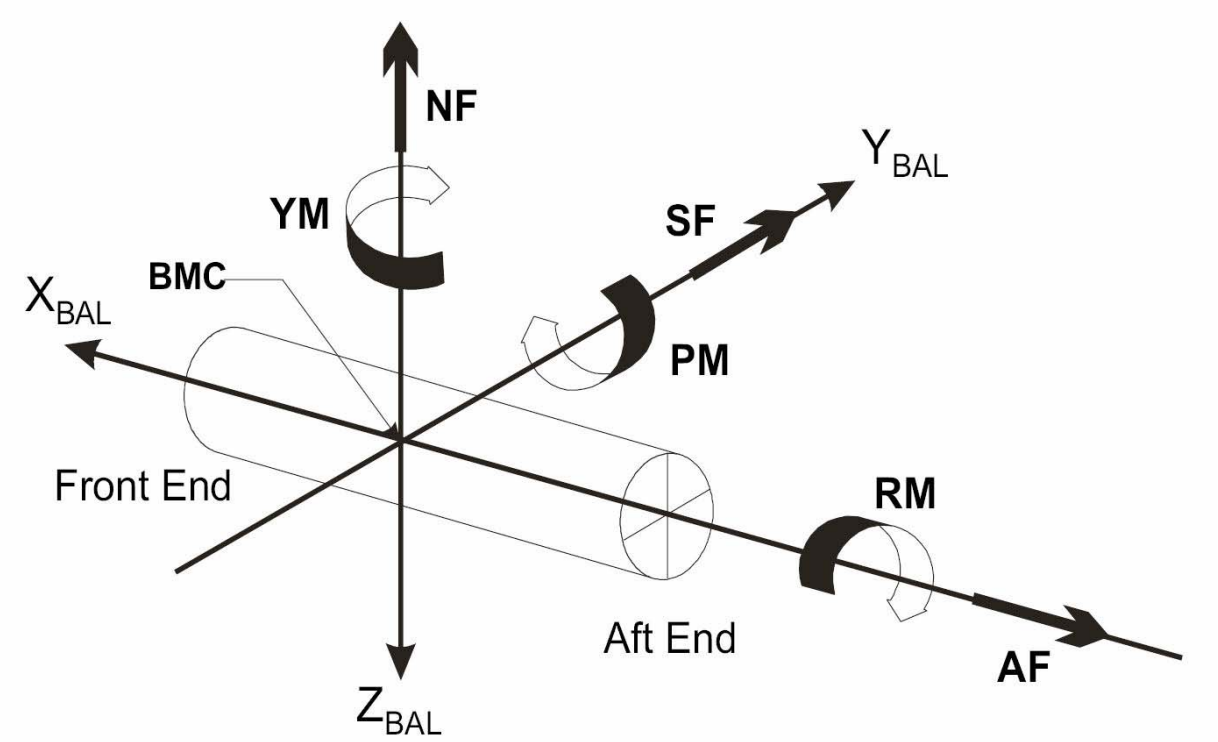

Figure 31. Coordinate System for Balance Forces and Moments Relative to Origin at the Balance Moment Center, BMC.

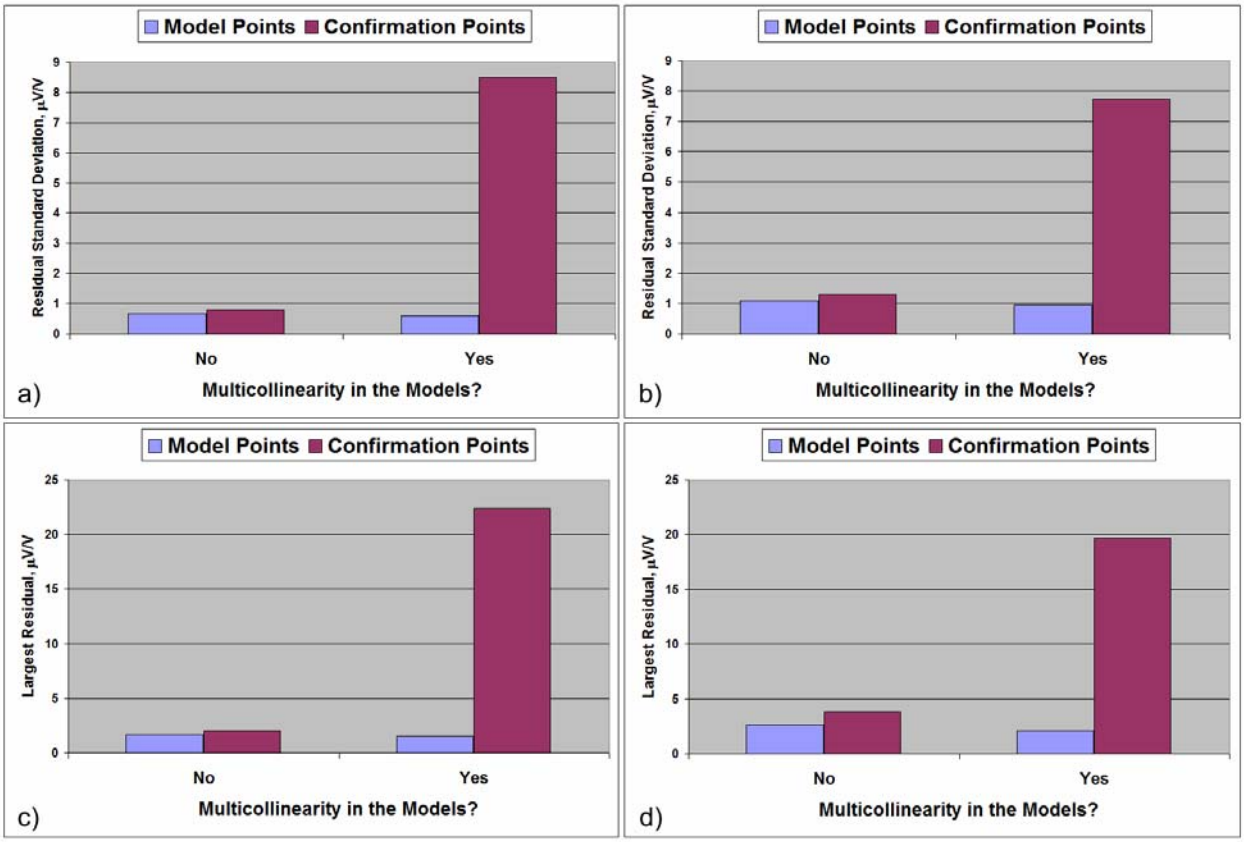

Figure 32. Effect of Multicollinearity on Model Residuals and Confirmation-Point Residuals. a) Residual Standard Deviation, Pitching Moment; b) Residual Standard Deviation, Yawing Moment; c) Largest Residual, Pitching Moment; d) Largest Residual, Yawing Moment. Multicollinearity affects confirmation-point predictions more than model-point predictions.